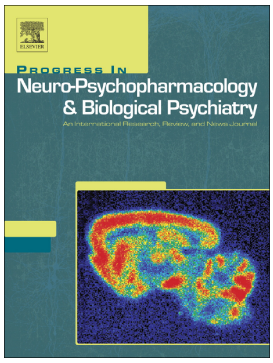
Accepted Manuscript



Application of machine learning classification for structural brain MRI in mood disorders: Critical review from a clinical perspective

Yong-Ku Kim, Kyoung-Sae Na

PII: S0278-5846(17)30213-0
DOI: doi: 10.1016/j.pnpbp.2017.06.024
Reference: PNP 9146

To appear in: *Progress in Neuropsychopharmacology & Biological Psychiatry*

Received date: 15 March 2017
Revised date: 7 June 2017
Accepted date: 22 June 2017

Please cite this article as: Yong-Ku Kim, Kyoung-Sae Na , Application of machine learning classification for structural brain MRI in mood disorders: Critical review from a clinical perspective. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Pnp(2017), doi: 10.1016/j.pnpbp.2017.06.024

Application of machine learning classification for structural brain MRI in mood disorders: Critical review from a clinical perspective

Yong-Ku Kim[1], Kyoung-Sae Na[2*]

[1]Department of Psychiatry, College of Medicine, Korea University, Seoul, Republic of Korea

[2]Department of Psychiatry, Gachon University Gil Medical Center, Incheon, Republic of Korea

Running title: Machine learning in mood disorders

Number

Word counts: 271 (abstract), 5233 (body), 2 Tables, 5 Figures

Corresponding authors

Kyoung-Sae Na, MD. Department of Psychiatry, Gachon University, Gil Medical Center, 21, 774beongil, Namdong-daero, Incheon 21565, Korea. Tel: +82-32-468-9932, Fax: +82-32-468-9962, Email: ksna13@gmail.com

Abstract

Mood disorders are a highly prevalent group of mental disorders causing substantial socioeconomic burden. There are various methodological approaches for identifying the underlying mechanisms of the etiology, symptomatology, and therapeutics of mood disorders; however, neuroimaging studies have provided the most direct evidence for mood disorder neural substrates by visualizing the brains of living individuals. The prefrontal cortex, hippocampus, amygdala, thalamus, ventral striatum, and corpus callosum are associated with depression and bipolar disorder. Identifying the distinct and common contributions of these anatomical regions to depression and bipolar disorder have broadened and deepened our understanding of mood disorders. However, the extent to which neuroimaging research findings contribute to clinical practice in the real-world setting is unclear. As traditional or non-machine learning MRI studies have analyzed group-level differences, it is not possible to directly translate findings from research to clinical practice; the knowledge gained pertains to the disorder, but not to individuals. On the other hand, a machine learning approach makes it possible to provide individual-level classifications. For the past two decades, many studies have reported on the classification accuracy of machine learning-based neuroimaging studies from the perspective of diagnosis and treatment response. However, for the application of a machine learning-based brain MRI approach in real world clinical settings, several major issues should be considered. Secondary changes due to illness duration and medication, clinical subtypes and heterogeneity, comorbidities, and cost-effectiveness restrict the generalization of the current machine learning findings. Sophisticated classification of clinical and diagnostic subtypes is needed. Additionally, as the approach is inevitably limited by sample size, multi-site participation and data-sharing are needed in the future.

Kew words: machine learning, neuroimaging, MRI, depression, bipolar disorder

Abbreviation

bipolar disorder, BD; major depressive disorder, MDD; magnetic resonance imaging (MRI); structural MRI, sMRI; functional MRI, fMRI; support vector machine, SVM; Principal component analysis, PCA; regions of interest, ROI; relevance vector machine, RVM; Gaussian Process Classification, GPC; selective serotonin reuptake inhibitors, SSRI; BD type I, BDI; BD type II, BDII; Sequenced Treatment Alternatives to Relieve Depression, STAR*D

Contents

1. Introduction

Mood disorder is a highly prevalent psychiatric disorder, with a higher prevalence in developed countries than in middle- to low-income countries (Bromet et al., 2011, Merikangas et al., 2011). For major depressive episodes, which commonly occur during the course of bipolar disorder (BD) and major depressive disorder (MDD), the average lifetime prevalence was 14.6% and 11.1% in high income and low- to middle-income countries, respectively. The economic burden of mood disorder shows that socioeconomic development and time-related progress does not guarantee disease control (Greenberg et al., 2015, Murray et al., 2012).

The dissociation between socioeconomic development and mood disorder raises the question of contributing factors. Generally, genetic, neuroendocrine, psychosocial, and neuroanatomical factors contribute to mood disorder development, progression, and treatment response (Kim et al., 2016, Na et al., 2014, Na et al., 2016). As the contributing factors and underlying mechanisms are complicated, no single cause can explain the etiology of a mood disorder; however, measuring the structure and function of the human brain can provide the most direct information for understanding mood disorders, as most mental and behavioral symptoms arise from neural structures and activity in the brain.

Numerous studies have investigated the neural substrates of mood disorders. Unfortunately, most results have not been translated to clinical practice, such as diagnosis and treatment. As the initial attention to neuroimaging studies was focused on identifying disease-specific mechanisms, the limitations of non-applicability were not a large pitfall. In the past decades, numerous neuroimaging studies have been conducted to identify factors that modulate the etiology, symptomatic characteristics, and treatment responses in BD and MDD. However,

critics have argued that traditional neuroimaging studies are limited in terms of clinical practice. (Phillips and Swartz, 2014). The continuing improvement of neuroimaging technology and numerous findings from these neuroimaging modalities should ideally contribute more directly to clinical practice (Stringaris, 2015). However, these advances do not necessarily guarantee pragmatic clinical solutions (Paulus, 2015).

Given its pragmatism, structural MRI (sMRI) is the most feasible method available for contributing to clinical practice. Positron emission tomography and single-photon emission computed tomography visualize receptor-specific function, especially accompanied by drugs (Heiss and Herholz, 2006). However, a relatively long measuring time, low anatomical resolution, and exposure to radioactive probes hinder their routine use in clinical practice. On the contrary, sMRI has high anatomical resolution (Erhart et al., 2005). Although functional MRI (fMRI) visualizes real-time brain activity by indirectly measuring regional blood flow (Glover, 2011), the long measuring time and the effort required to perform a specific task would be difficult for patients suffering from a mood disorder. Hence, given the pragmatic utility of sMRI, this paper reviews the application of machine learning to neuroimaging in mood disorders from the view of a researcher and physician.

2. Current outcomes and limitations of sMRI

2.1. Summaries of current findings

Recent well-organized meta-analyses provide a detailed description of structural neuroimaging findings for MDD and BD (Niu et al., 2017, Wise et al., 2014, Wise et al., 2016b). Thus, it is out of the scope of this review to comprehensively describe results from

neuroimaging studies in MDD and BD. In this section, we briefly summarize the common and distinct findings of sMRI studies of MDD and BD.

The common and distinct regions of gray matter in MDD and BD were described in a recent meta-analysis (Wise et al., 2016b). The commonly reduced gray matter regions in both MDD and BD include bilateral insula, right superior temporal gyrus, bilateral anterior cingulate cortex, and left superior medial frontal cortex. On the other hand, the right cerebellar vermis, right middle frontal gyrus, left inferior parietal lobule, and left hippocampus were significantly decreased in MDD compared to BD.

Regarding white matter regions, the genu of the corpus callosum, which extends to the left prefrontal white matter, was decreased in both MDD and BD compared to healthy controls. BD patients had significantly decreased fractional anisotropy values in the posterior cingulum compared to MDD patients (Wise et al., 2016a). The predominantly decresed gray matter in MDD and white matter in BD suggest differential brain structure roles in MDD and BD.

Cortical thickness was measured later than gray matter volumes, which were mainly measured by voxel-based morphometry (VBM). Cortical thickness generally represents the shortest distance between the boundary of gray/white matter and the pial surface (Dale et al., 1999, Rakic, 2008), whereas gray matter volume represents the combination of two genetically independent anatomical properties: cortical surface area and thickness (Winkler et al., 2010). A recent meta-analysis showed that the rostral middle frontal cortex was thinned in BD compared to MDD (Niu et al., 2017). The commonly thinned cortical regions among MDD and BD included the left inferior temporal cortex and left fusiform gyrus.

2.2. Limitations

As briefly described in the above section, we now know which regions or tracts are preferentially or commonly associated with MDD and BD. This helps us understand the key neuroanatomical features and their relationships with other important contributing factors such as genetics, neuroendocrine issues, and psychosocial stress.

Despite findings from numerous studies and meta-analytic summaries, we are still uncertain if an individual has MDD, BD, or is psychiatrically 'normal' based on his/her neuroimaging scan. The discrepancy between research and clinical practice is mainly attributable to the basic statistical approach and object of the statistics. Traditional neuroimaging studies are based on group-level statistics, which aim to identify differences at the group level. The hippocampus is a representative anatomical region in which MDD patients had decreased volumes compared to healthy controls. However, if an individual with MDD has a similar hippocampal volume to healthy controls, how we can make a diagnosis based on sMRI? (**Figure 1**) The knowledge that hippocampal volume is decreased in MDD is only useful understanding neuroanatomical characteristics of MDD, but not individual patients with MDD.

To provide a solution to this problem, an individual-level statistical approach is required for neuroimaging studies. A machine learning-based approach is the most feasible and widely investigated method of individual-level comparisons. In the next section, we begin in earnest to review machine learning and its application in neuroimaging studies. Since several previous reviews focused on the methodological aspects of machine learning in neuroimaging, we have attempted to focus on machine learning from the perspective of clinical psychiatrists.

3. Machine learning

3.1. Definition and history

Arthur Samuel first used the term "machine learning" (Samuel, 1959). The definition of machine learning was the "field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959). The most relevant definition of machine learning in the field of neuroscience was coined by Tom Mitchell. According to Mitchell, a computer program can be said to learn from experience if a performance at a specific task is improved with experience (Mitchell, 1997). From a methodological perspective, machine learning consists of representation, evaluation, and optimization (Domingos, 2012). These three steps are essential to the machine learning approach for neuroimaging studies. As most machine learning algorithms handle a substantial number of variables, machine learning shares many features with big data science. Machine learning is developing quickly and produces robust outcomes in various areas, including medicine (Marr, 2016). For example, IBM Watson for Oncology has been increasingly disseminated in the clinical setting (Shrager and Tenenbaum, 2014).

3.2. Why machine learning in neuroimaging?

The advantage of machine learning is not confined to predicting the future or unknown objects. Several non-machine learning statistical methods such as regression analysis have been used to predict the future given past and present data. Rather, the most salient advantage of machine learning is its applicability for individual-level analysis. Most traditional statistical methods compute group level significance and effects. The numerous studies

included in the meta-analyses mentioned in Section 2.1. are typical examples of group-level findings using neuroimaging modalities. Due to the results from the group-level analyses of neuroimaging studies, now neuroscientists and physicians can better understand the underlying neural mechanisms of MDD and BD. However, this does not necessarily guarantee proportional aid in the diagnosis and treatment of individual patients with MDD and BD. Machine learning addresses this point, and increasing attention and trials have examined the utility of machine learning in the clinical field.

3.3. Types of machine learning

Machine learning can be categorized into three types: supervised learning, unsupervised learning, and reinforcement learning.

3.3.1. Supervised learning

The essential of machine learning is to create a computer algorithm which can distinguish patients with MDD from healthy people. Once the algorithm is made, then training is required to learn how to discriminate patients with mood disorders versus healthy individuals. Thus, neuroimaging data from all participants should be labelled as diseased or healthy. The clinical classification should be conducted by experienced mental health professionals such as psychiatrists, although it is not possible to perfectly classify as described in Section 4.2.2.

This type of learning, which requires pre-determined information from outside the learning algorithm is called supervised learning because the labelling is externally created. The most common type of supervised learning in neuroimaging analysis is a support vector machine (SVM) (see Section 3.5.1.).

3.3.2. Unsupervised learning

Unsupervised learning refers to a process by which a computer algorithm conducts analyses without any externally defined information. Principal component analysis (PCA) is a frequently used type of unsupervised learning in neuroimaging (see Section 3.4.2.).

3.3.3. Reinforcement learning

A key concept of reinforcement learning is interaction with the environment. Involvement of external factors is like supervised learning. However, unlike the fixed label in the supervised learning, the feedback from the environment is flexible per the strategy of the machine. Deep learning is a well-known type of reinforcement learning. Reinforcement learning produced substantial outcomes in games such as Go (Gibney, 2016). However, there has not yet been a promising outcome or utility for reinforcement learning in neuroimaging.

3.4.Major methodological issues

3.4.1. Curse of dimensionality

As the phrase 'curse of dimensionality' (Bellman, 2015) implies, this is a very common and serious problem in the field of machine learning. As the dimensionality increases, larger sample sizes are required to optimize algorithms. However, it is difficult to acquire a sufficient number of subjects in neuroimaging studies. The small sample sizes accompanied by an inevitably substantially higher number of dimensions results in over-fitting, which is

frequently present in neuroimaging machine learning studies. sMRI data in the training set contains both useful information and noise. As more dimensionalities are included in the building algorithm, the final version of the classification model is literally over-fitted to the characteristics in the training set. The attenuated generalizability is a fatal flaw and mostly cannot be applied in practical settings. Besides dimensionality reduction, standardization with parameter C (see Section 3.5.1.) is another important strategy to avoid over-fitting.

3.4.2. Feature selection

To avoid the curse of dimensionality, various levels of approach are needed. Feature selection is the first step to decrease unnecessary information. Raw materials contain every sort of information. Generally, it is suggested that 10^6 brain dimensions are created when analyzing sMRI with VBM (Keogh and Mueen, 2010). Thus, not all the brain structures are essential to differentiate depressed patients from healthy controls. If one can select essential brain structures and discard needless one, like bone a fish, subsequent analysis could be done with more appropriate materials.

Type of feature selection techniques are comprehensively described in a previous review (Saeys et al., 2007). In a nutshell, two types of feature selection could be conducted. One type selects features within whole data, while the other type selectively extracts regions of interest (ROI) based on a hypothesis or findings from previous studies. The former method is again divided into two methods. The first utilizes unsupervised machine learning methods such as PCA, while the second uses a simple univariate t-test at each region to select statistically different regions.

3.4.2.1. PCA

The primary function of PCA is to reduce a larger set of variables into a smaller set of 'artificial' variables (called principal components) that account for most of the variance in the original variables (Kaiser, 1960). Variables are supposed to be continuous and have linear relationships (Shlens, 2014), and PCA reduces dimensionality based on the relationships among variables. **Figure 2** shows a simplified view of PCA dimensionality reduction.

The major strength of PCA is that it is based on whole brain voxel data. Although meta-analyses have identified major neuroanatomical regions and pathways which are closely associated with MDD and BD (see Section 2.1.), most anatomical regions including the brain stem (Steele et al., 2005) and cerebellum (Zhao et al., 2016) have been reported to be associated with mood disorders. The concept of connectivity also emphasizes connections and interplay among various regions rather than specific anatomical regions (Jiang et al., 2016, Nortje et al., 2013). Thus, if dimensionality reduction is conducted upon the whole brain, the complicated connectivity may be fully reflected during dimensionality reduction. Several machine learning studies have used PCA to reduce dimensionality (Fu et al., 2008).

3.4.2.2. Univariate independent *t*-test

The univariate t-test is a simple statistical method. First, each region for two groups is compared by univariate *t*-test. Then, regions showing statistically significant differences between the two groups can be selected in the algorithm and optimization stages. The major limitation of this method is that the statistically significant differences between regions obtained from the group-level analysis do not guarantee that these regions deserved to be selected (see Section 3.3.2.) The major strength of the univariate t-test is the simplicity.

Hence, despite the flaws, feature selections with univariate t-tests have been widely used in several neuroimaging machine learning studies (Mwangi et al., 2012).

3.4.2.3.Masking ROI

The masking ROI approach is distinct from the former two-feature selection methods, the PCA and univariate *t*-test. Unlike the previous two methods, masking ROI excludes several anatomical or connectivity regions based on an *a priori* hypothesis or accumulated evidence. One of the most commonly used means for creating software ROI masks is the WFU Pickatlas. (Version 3.03; http://www.fmri.wfubmc.edu). The cons of the masking ROI are exactly the pros of the whole-brain approach. The masking ROI approach excludes regions which could possibly be associated with those that properly classify mood disorder.

The masking ROI was shown to be a good feature selection method when there was a small sample size (Chu et al., 2012). Although neurocognitive disorders such as mild cognitive impairment and Alzheimer's disease were the target in that study, we briefly introduce several findings from that study here. In that study, there were no differences in classification accuracy between a whole-brain data-driven approach and a masking ROI method when discriminating Alzheimer's disease patients from healthy controls. However, in the discrimination between MCI and healthy controls, the masking ROI method showed better performance with a small sample size, while the whole brain approach was better with a large sample size (n = 200). Interestingly, even with a sample size of 200, the whole brain approach was not better than the ROI approach with respect to the hippocampus as well as the parahippocampal gyrus (Chu et al., 2012). Whole brain approaches did not show better classification accuracy with a sample size of 60 or less when compared to any configuration of brain regions in the ROI. In the field of mental and behavioral disorders, several machine

learning studies used ROI approaches and produced good classification accuracy (Grotegerd et al., 2013).

3.5.Classifier

Based on the selected features, a computer algorithm can be trained to optimize its classifying model. The most commonly used classifiers include SMV. Recently, Bayesian model-based classifiers such as relevance vector machine (RVM) and Gaussian Process Classification (GPC) have also been used.

3.5.1. SVM

The SVM has been used to solve clinical problems since the mid-1990s (Cortes and Vapnik, 1995). SVM divides data (or a sample) into two groups using a linear classifier **(Figure 3)**.

If a classifier is too close to any side of the two groups, then the possibility of misclassification increases. Hence, it is needed for the classifier to secure a maximum margin from each sample. Two lines parallel to the classifier should be drawn along each side. Data points of each group nearest a classifier are called support vectors. Support vectors lie on each boundary of an optimal hyperplane. An optimal line (red line in Figure 7) can classify new data points only with support vectors, but not with entire data points.

The original version of SVM utilized the hard margin of classification boundaries. The 'hard margin' means no allowance of violating the line. Unfortunately, there is no perfect classifier As in **Figure 3**, if one sample is above the upper decision boundary, then it would be classified as Group A. However, if a classifier has to perfectly discriminate all variables in a

training set, then the margin could be too narrow, which results in over-fitting and might inappropriately classify new cases. To make up for this weak point, the concept of parameter C was proposed (Cortes and Vapnik, 1995). As the C value is set larger, both the margin and the error rates increase, which contribute to high bias and low variance. By adjusting the value of the C parameter, classification can be balanced appropriately (**Figure 4**). Several SVM libraries set the default value of C as 1, which is adjustable according to the objectives and properties of each sample.

As most SVM data does not have a linear distribution, the SVM uses a kernel function (also known as the kernel trick) to convert complicated multiple dimensions into a space where the linear hyperplane can function.

3.5.2. Bayesian model

The Bayesian model is named after Thomas Bayes (1701-1761). The fundamental assumption of the Bayesian model is that the probability of a prior event influences the determination of its posterior probability (Stigler, 1982). Unlike the non-Bayesian model (also known as frequentist) which assumes a fixed parameter, the Bayesian method uses a probabilistic parameter ranging from 0 to 1, which is also called the posterior distribution.

GPC and RVM have been used in neuroimaging studies. RVM shares a key learning algorithm assumption with SVM (Tipping, 2001). One of the major differences is a probabilistic outcome ranging from 0 to 1. This probabilistic outcome is sometime useful for the adjustment of predictive uncertainty, such as disease severity and subtypes (Bishop, 2006).

A recent study revealed that bipolar patients with a higher number of manic episodes were classified into a higher certainty group, whereas those with a lower number of manic episodes

belonged to the lower certainty group (Mwangi et al., 2016).

Given the theoretical and practical advantages, RVM has been increasingly used along with SVM to classify mood disorders (Passos et al., 2016).

4. Major issues of machine learning in clinical practice

Generally, neuroimaging studies for which machine learning is applied are divided into three categories: (1) prediction of disease onset or progression, (2) diagnosis, and (3) prediction of treatment response. In the clinical setting, a machine learning approach serves two ends. The sMRI taken at baseline can be used for diagnostic classification and for predicting treatment response.

4.1. Prediction of disease onset or progression

Predicting the risk of disease using machine learning has been developed in various fields of medicine (Burki, 2016, Tripoliti et al., 2017). The most widely investigated psychiatric disease may be Alzheimer's disease. Several studies made machine learning models to predict the onset of Alzheimer's disease or the conversion from MCI to Alzheimer's disease (Wei et al., 2016, Young et al., 2013).

However, in the case of MDD and BD, those predictive approaches could be limited by several issues. First, the importance of non-predictable factors cannot be controlled. Psychiatric disorders do not develop solely from neurobiological substrates. Rather, most psychiatric disorders arise from the interaction of the environment with genetics. In particular, MDD has a particularly low heritability rate (0.3 to 0.4) (Kendler and Aggen, 2001, Wray and

Gottesman, 2012), so it is difficult to predict the future onset of MDD given current evidence of associated neural substrates.

However, efforts to apply machine learning to diagnosis and treatment in mood disorder have been increasing. A recent prospective 5-year follow-up study with a machine learning approach was conducted (Foland-Ross et al., 2015). That study was based on adolescent girls and reported that cortical thickness could predict differences between subjects who later experienced depression and subjects who remained healthy. The predictive accuracy for the onset of MDD was about 70% ($p = 0.021$).

4.2. Diagnosis

Diagnostic classification is the most widely and frequently applied area of machine learning studies in neuroimaging. Generally, the diagnostic accuracy of machine learning studies in mood disorders ranged from 67.6% (Costafreda et al., 2009) to 90.3% (Mwangi et al., 2012). A recent literature review described machine learning studies in depression from a methodologic perspective (Patel et al., 2016).

4.2.1. Heterogeneity within diagnosis

In mood disorders such as MDD and BD, there are various subtypes within the same diagnosis. Major clinical variables such as clinical symptoms, prognosis, and treatment response differ between subtypes. For example, response to selective serotonin reuptake inhibitors (SSRI) is relatively low in patients with prominent anhedonia (Uher et al., 2012).

Although major neuroanatomical regions and neural networks are common in one

diagnostic entity, prominently disturbed features differ between subtypes. Melancholic depression, which is commonly accompanied by anhedonia and psychomotor retardation, is associated with the medical forebrain bundle in MDD (Bracht et al., 2014). On the other hand, rumination and autobiographical memories were commonly associated with abnormal connectivity in the default mode network (Lois and Wessa, 2016). Within BD, differences in neuroanatomical structures according to subtype were not consistent. Although a recent study showed that BD type I (BDI) had substantially decreased cortical volume and thickness compared to BD type II (BDII) (Abe et al., 2016), few studies have investigated structural abnormalities according to BD type (Phillips and Swartz, 2014). In addition, findings from studies suggesting differences between BDI and BDII were inconsistent (Ha et al., 2011, Liu et al., 2010).

4.2.2. Diagnostic uncertainty

Except in forensic or consultation-liaison psychiatry, in which malingering or compensation-related issues hinder a valid diagnosis, most clinicians may not have difficulty judging whether a patient was psychiatrically ill or healthy. However, clinicians may want to receive support in diagnosing patients with a first episode, and many differential diagnostic considerations are needed. Additionally, clinicians may want to differentiate diagnoses which are not easily distinguished from each other by clinical variables.

Valid diagnosis requires longitudinal follow-up after the initial evaluation. The first depressed episode does not guarantee that the person will have a life-long diagnosis of MDD (Angst et al., 2011). Diagnosing psychosis is more complicated if it is only based on a cross-sectional evaluation. Individuals with psychotic features could have MDD, BD, schizophrenia, or other psychotic disorders. No matter how good clinicians are at psychiatric

diagnosis, it is impossible to predict the future of those diseases. One research group followed-up patients with a psychotic mood disorder for 10-years (Ruggero et al., 2010, Ruggero et al., 2011). During the diagnostic evaluation points at 6 months, 2 years, and 10 years, 97 (49.7%) out of 195 patients with BD had one psychiatric disorder other than BD (Ruggero et al., 2010). In that study, the proportion of patients who had been stably diagnosed with BD for the 10-year period was only 50.3%. Another study based on 146 patients with MDD by the same group showed more instability (Ruggero et al., 2011). Only 37.7% (55 out of 146) of patients were diagnosed with MDD and had not been switched to another diagnosis. In a retrospective review of medical charts, 46 out of 250 (18.4%) patients with MDD shifted to BD within 5 years (Woo et al., 2015). When 409 patients who had been diagnosed with recurrent MDD were reassessed, 40.8% (167 out of 409) were diagnosed with BDII (Mosolov et al., 2014).

Those results suggested that it is necessary to conduct longitudinal studies to obtain diagnostically validated samples which are guaranteed by clinicians as well as time. However, many studies recruited chronic and/or recurrent patients for the optimization of the machine learning-aided classification **(Table 1)**.

4.3. Predicting treatment response

Predicting individual treatment response is substantially important for both physicians and patients. In particular, the efficacy of antidepressants for the treatment of MDD has been a major focus of clinical practice. In the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (Rush et al., 2006), the cumulative remission rate within a year was only 67%.

Only a few machine learning studies have been conducted to predict treatment response. The low response to antidepressant treatment in clinical practice raised a need for precise prediction of optimized treatment for mood disorders. In one study, accuracy of the discriminating treatment response and resistance by machine learning was 82.9% (Liu et al., 2012).

5. Current limitations and future directions

5.1. Effects of pharmacotherapy

Pharmacological and non-pharmacological treatments exert their therapeutic effects by modulating neuromolecular systems (Barsaglini et al., 2014). Pharmacotherapy contributed to a broader range of large changes in the brain when compared to psychotherapy (Boccia et al., 2016). In that study, changes in the anterior cingulate and striatum were particularly large. Although there were some inconsistencies, larger volumes in the hippocampus, dorsolateral prefrontal cortex, and medial prefrontal cortex have been considered prognostic factors for better antidepressant response (Bellani et al., 2011, Dusi et al., 2015). Antidepressants can increase the volume of the above regions. Recent studies also reported changes in occipital regions, which were not included in the essential neuroanatomical regions in which antidepressants exert their effects (Jung et al., 2014). Possible mechanisms by which antidepressants and lithium lead to volume increases include reversal of neuronal and glial cell death and synaptic plasticity (Banasr et al., 2011, Gray and McEwen, 2013, Kim et al., 2016). In BD, lithium is associated with increased white matter integrity (Gildengers et al., 2015) and hippocampal volumes (Hajek et al., 2014, Hajek et al., 2012).

Despite the importance of antidepressants and lithium for neuromolecular and anatomical alterations in the brain, many machine learning studies included patients on medication.

5.2. Age at onset, duration of illness, and recurrence

Patients who had a mood disorder decades ago and have been taking medications for a long period of time tend to have a well-known diagnosis and an optimized treatment strategy. However, the long-time is a double-edged sword. Age at onset, duration of illness, and recurrence are all basic clinical variables which contribute to structural changes in the brain. Many studies have consistently suggested that duration of illness and recurrence substantially influenced structural brain abnormalities, aside from the possible effects of medication (Na et al., 2016).

On the other hand, a longer duration of illness was associated with lower FA in MDD (de Diego-Adelino et al., 2014). The duration of illness was negatively correlated with superior cerebellar peduncular volume in MDD (Zhao et al., 2016).

Regarding the age of onset, amygdala volume was decreased in early onset disease and with disease progression (Usher et al., 2010). A decreased genu of the corpus callosum was particularly associated with early onset MDD (Kemp et al., 2013). Late-onset depression was associated with less hippocampal atrophy than early-onset MDD (Lloyd et al., 2004). In a recent study on BD, age at onset was associated with inferior frontal gyrus cortical thinning (Knochel et al., 2016).

5.3. Limitations in the neuroimaging machine learning

Careful selection and recruitment of clinical samples is essential for an optimized model

which has the ability to diagnostically classify individuals. However, many machine learning studies did not fully consider the importance of clinical variables **(Table 2)**. Some studies included chronic patients with a mean age of 51.8 years (Johnston et al., 2015). Most studies were conducted with a cross-sectional design except one study with a 1-year follow-up (Serpa et al., 2014). Major clinical variables were not measured or reported in most articles (Gong et al., 2011, Jie et al., 2015, Mwangi et al., 2012, Schnack et al., 2014).

In the field of mood disorder, this is the time to focus on the importance of clinical variables as well as the learning algorithms. To overcome the limitations of machine learning studies, sufficient sample size and time are necessary. Given the heterogeneities in diagnoses, diagnostic uncertainty due to cross-sectional designs, and the possible effects of medications on structural brain regions, longitudinal follow-up is also needed.

It might not be possible for individual investigators to acquire sufficient sample sizes or have the time to recruit a homogeneous sample for their machine learning training set. The current solution is to collect and share data together. Recently, cortical abnormalities in adolescents and adults with MDD were examined in data from the multi-site large-scale consortium ENIGMA study (Enhancing Neuro Imaging Genetics through Meta-Analysis) (Schmaal et al., 2016, Thompson et al., 2017). Data sharing can be also helpful for neuroimaging machine learning studies (Poldrack and Gorgolewski, 2014). Several institutions such as the INCF Task Force on Neuroimaging Datasharing (Poline et al., 2012) and The Neuroimaging Informatics Tools and Resources Clearinghouse (Kennedy et al., 2016) have proceeded to share neuroimaging data for researchers.

5.4. Why should neuroimaging be used for machine learning?

Although structural MRI is feasible for machine learning studies which are cost-effective and have high-resolution images, it is more still expensive than utilizing clinical variables or peripheral blood. Thus, if such non-neuroimaging modalities perform diagnostic classification with reasonable accuracy, why is a machine learning approach with sMRI needed? For a traditional, univariate, group-level analysis, the neuroimaging modality can provide neural substrates for the disorder. However, if only accuracy is important and other disease-specific information is not under consideration, it is unclear why clinicians should use sMRI for machine learning-aided support.

MDDScore, which consists of 9 serum biomarkers such as brain-derived neurotrophic factor, epidermal growth factor, α1 antitrypsin, apolipoprotein C3, soluble tumor necrosis factor α receptor type 2, myeloperoxidase, resistin, and prolactin, has an overall diagnostic accuracy of 91% (Bilello et al., 2015). In another study using nuclear magnetic resonance-based plasma metabolomics and SVM, the accuracy was 88% (Zheng et al., 2017). Non-neuroimaging machine learning has been widely investigated for predicting treatment response. A recent study analyzed results from the STAR*D trial (Chekroud et al., 2016). In that study, the accuracy for the predictability of treatment response was 51-65%. An electroencephalogram-based machine learning study reported the predictive accuracy of SSRI treatment response as 87.9% (Khodayari-Rostamabad et al., 2013).

The strength of neuroimaging over peripheral biomarkers is that it is not significantly influenced by transient systemic conditions such as overt inflammation, nutritional status, and endocrinologic disturbances. Although the sMRI results could be also influenced by all the systemic conditions including inflammatory, nutritional, and endocrinologic factors, the accompanied changes are not as temporal as peripheral markers which are altered even within

hours. Thus, neuroimaging modalities may be superior to peripheral blood-based approaches.

5.5. Future direction: integration of data with different scales

The process of machine learning in neuroimaging for mood disorder is summarized in **Figure 5**. However, can a single brain sMRI give us a solution to the diagnosis of mood disorder and treatment response to medications in the near future? Neuroimaging machine learning, however good, can only provide a piece of the puzzle. Clinical data such as current severity of depression, age at onset, personality traits, early-life experience, current environmental factors, substance behavior such as alcohol and smoking, general medical conditions could not be all included in a sMRI. As there is no way to reliably integrate all the variables with different scales into one model of machine learning, further improvement in machine learning algorithms and optimization is warranted.

Acknowledgments

There is no funding source for this paper.

All authors declare that there is no conflict of interest.

REFERENCES

Abe C, Ekman CJ, Sellgren C, Petrovic P, Ingvar M, Landen M. Cortical thickness, volume and surface area in patients with bipolar disorder types I and II. J Psychiatry Neurosci. 2016;41:240-50.

Angst J, Azorin JM, Bowden CL, Perugi G, Vieta E, Gamma A, et al. Prevalence and characteristics of undiagnosed bipolar disorders in patients with a major depressive episode: the BRIDGE study. Arch Gen Psychiatry. 2011;68:791-8.

Banasr M, Dwyer JM, Duman RS. Cell atrophy and loss in depression: reversal by antidepressant treatment. Curr Opin Cell Biol. 2011;23:730-7.

Barsaglini A, Sartori G, Benetti S, Pettersson-Yeo W, Mechelli A. The effects of psychotherapy on brain function: a systematic and critical review. Prog Neurobiol. 2014;114:1-14.

Bellani M, Dusi N, Yeh PH, Soares JC, Brambilla P. The effects of antidepressants on human brain as detected by imaging studies. Focus on major depression. Prog Neuropsychopharmacol Biol Psychiatry. 2011;35:1544-52.

Bellman RE. Adaptive control processes: a guided tour: Princeton university press; 2015.

Bilello JA, Thurmond LM, Smith KM, Pi B, Rubin R, Wright SM, et al. MDDScore: confirmation of a blood test to aid in the diagnosis of major depressive disorder. J Clin Psychiatry. 2015;76:e199-206.

Bishop C. Pattern Recognition and Machine Learning. New York: Springer-Verlag New York; 2006.

Boccia M, Piccardi L, Guariglia P. How treatment affects the brain: meta-analysis evidence of neural substrates underpinning drug therapy and psychotherapy in major depression. Brain Imaging Behav. 2016;10:619-27.

Bracht T, Horn H, Strik W, Federspiel A, Schnell S, Hofle O, et al. White matter microstructure alterations of the medial forebrain bundle in melancholic depression. J Affect Disord. 2014;155:186-93.

Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, de Girolamo G, et al. Cross-national epidemiology of DSM-IV major depressive episode. BMC Med. 2011;9:90.

Burki TK. Predicting lung cancer prognosis using machine learning. Lancet Oncol. 2016;17:e421.

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3:243-50.

Chu C, Hsu AL, Chou KH, Bandettini P, Lin C, Alzheimer's Disease Neuroimaging I. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage. 2012;60:59-70.

Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20:273-97.

Costafreda SG, Chu C, Ashburner J, Fu CH. Prognostic and diagnostic potential of the structural neuroanatomy of depression. PLoS One. 2009;4:e6353.

Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage. 1999;9:179-94.

de Diego-Adelino J, Pires P, Gomez-Anson B, Serra-Blasco M, Vives-Gilabert Y, Puigdemont D, et al. Microstructural white-matter abnormalities associated with treatment resistance, severity and duration of illness in major depression. Psychol Med. 2014;44:1171-82.

Domingos P. A few useful things to know about machine learning. Commun ACM. 2012;55:78-87.

Dusi N, Barlati S, Vita A, Brambilla P. Brain Structural Effects of Antidepressant Treatment in Major Depression. Curr Neuropharmacol. 2015;13:458-65.

Erhart SM, Young AS, Marder SR, Mintz J. Clinical utility of magnetic resonance imaging radiographs for suspected organic syndromes in adult psychiatry. J Clin Psychiatry. 2005;66:968-73.

Foland-Ross LC, Sacchet MD, Prasad G, Gilbert B, Thompson PM, Gotlib IH. Cortical thickness predicts the first onset of major depression in adolescence. Int J Dev Neurosci. 2015;46:125-31.

Fu CH, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SC, et al. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biological psychiatry. 2008;63:656-62.

Gibney E. Google AI algorithm masters ancient game of Go. Nature. 2016;529:445-6.

Gildengers AG, Butters MA, Aizenstein HJ, Marron MM, Emanuel J, Anderson SJ, et al. Longer lithium exposure is associated with better white matter integrity in older adults with bipolar disorder. Bipolar Disord. 2015;17:248-56.

Glover GH. Overview of functional magnetic resonance imaging. Neurosurg Clin N Am. 2011;22:133-9, vii.

Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, et al. Prognostic prediction of therapeutic response in depression using high-field MR imaging. Neuroimage. 2011;55:1497-503.

Gray JD, McEwen BS. Lithium's role in neural plasticity and its implications for mood disorders. Acta Psychiatr Scand. 2013;128:347-61.

Greenberg PE, Fournier AA, Sisitsky T, Pike CT, Kessler RC. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry. 2015;76:155-62.

Grotegerd D, Suslow T, Bauer J, Ohrmann P, Arolt V, Stuhrmann A, et al. Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: a pilot study. Eur Arch Psychiatry Clin Neurosci. 2013;263:119-31.

Ha TH, Her JY, Kim JH, Chang JS, Cho HS, Ha K. Similarities and differences of white matter connectivity and water diffusivity in bipolar I and II disorder. Neurosci Lett. 2011;505:150-4.

Hajek T, Bauer M, Simhandl C, Rybakowski J, O'Donovan C, Pfennig A, et al. Neuroprotective effect of lithium on hippocampal volumes in bipolar disorder independent of long-term treatment response. Psychol Med. 2014;44:507-17.

Hajek T, Kopecek M, Hoschl C, Alda M. Smaller hippocampal volumes in patients with bipolar disorder are masked by exposure to lithium: a meta-analysis. J Psychiatry Neurosci. 2012;37:333-43.

Heiss WD, Herholz K. Brain receptor imaging. J Nucl Med. 2006;47:302-12.

Jiang J, Zhao YJ, Hu XY, Du MY, Chen ZQ, Wu M, et al. Microstructural brain abnormalities in medication-free patients with major depressive disorder: a systematic review and meta-analysis of diffusion tensor imaging. J Psychiatry Neurosci. 2016;42:150341.

Jie NF, Zhu MH, Ma XY, Osuch EA, Wammes M, Theberge J, et al. Discriminating Bipolar Disorder From Major Depression Based on SVM-FoBa: Efficient Feature Selection With Multimodal Brain Imaging Data. IEEE Trans Auton Ment Dev. 2015;7:320-31.

Johnston BA, Steele JD, Tolomeo S, Christmas D, Matthews K. Structural MRI-Based Predictions in Patients with Treatment-Refractory Depression (TRD). PLoS One. 2015;10:e0132958.

Jung J, Kang J, Won E, Nam K, Lee MS, Tae WS, et al. Impact of lingual gyrus volume on antidepressant response and neurocognitive functions in Major Depressive Disorder: a voxel-based morphometry study. J Affect Disord. 2014;169:179-87.

Kaiser HF. The application of electronic computers to factor analysis. Educational and psychological measurement. 1960;20:141-51.

Kemp A, MacMaster FP, Jaworska N, Yang XR, Pradhan S, Mahnke D, et al. Age of onset and corpus callosal morphology in major depression. J Affect Disord. 2013;150:703-6.

Kendler KS, Aggen SH. Time, memory and the heritability of major depression. Psychol Med. 2001;31:923-8.

Kennedy DN, Haselgrove C, Riehl J, Preuss N, Buccigrossi R. The NITRC image repository. Neuroimage. 2016;124:1069-73.

Keogh E, Mueen A. Curse of Dimensionality. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer US; 2010. p. 257-8.

Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, Maccrimmon DJ. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. Clin Neurophysiol. 2013;124:1975-85.

Kim YK, Na KS, Myint AM, Leonard BE. The role of pro-inflammatory cytokines in neuroinflammation, neurogenesis and the neuroendocrine system in major depression. Prog Neuropsychopharmacol Biol Psychiatry. 2016;64:277-84.

Knochel C, Reuter J, Reinke B, Stablein M, Marbach K, Feddern R, et al. Cortical thinning in bipolar disorder and schizophrenia. Schizophr Res. 2016;172:78-85.

Liu F, Guo W, Yu D, Gao Q, Gao K, Xue Z, et al. Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS One. 2012;7:e40968.

Liu JX, Chen YS, Hsieh JC, Su TP, Yeh TC, Chen LF. Differences in white matter abnormalities between bipolar I and II disorders. J Affect Disord. 2010;127:309-15.

Lloyd AJ, Ferrier IN, Barber R, Gholkar A, Young AH, O'Brien JT. Hippocampal volume change in depression: late- and early-onset illness compared. The British journal of psychiatry : the journal of mental science. 2004;184:488-95.

Lois G, Wessa M. Differential association of default mode network connectivity and rumination in healthy individuals and remitted MDD patients. Soc Cogn Affect Neurosci. 2016;11:1792-801.

Marr B. A Short History of Machine Learning -- Every Manager Should Read. Forbes; 2016.

Merikangas KR, Jin R, He JP, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch Gen Psychiatry.

2011;68:241-51.

Mitchell TM. Machine learning. Boston, MA: WCB/McGraw-Hill; 1997.

Mosolov S, Ushkalova A, Kostukova E, Shafarenko A, Alfimov P, Kostyukova A, et al. Bipolar II disorder in patients with a current diagnosis of recurrent depression. Bipolar Disord. 2014;16:389-99.

Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380:2197-223.

Mwangi B, Ebmeier KP, Matthews K, Steele JD. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain. 2012;135:1508-21.

Mwangi B, Wu MJ, Cao B, Passos IC, Lavagnino L, Keser Z, et al. Individualized Prediction and Clinical Staging of Bipolar Disorders using Neuroanatomical Biomarkers. Biol Psychiatry Cogn Neurosci Neuroimaging. 2016;1:186-94.

Na KS, Lee KJ, Lee JS, Cho YS, Jung HY. Efficacy of adjunctive celecoxib treatment for patients with major depressive disorder: a meta-analysis. Prog Neuropsychopharmacol Biol Psychiatry. 2014;48:79-85.

Na KS, Won E, Kang J, Chang HS, Yoon HK, Tae WS, et al. Brain-derived neurotrophic factor promoter methylation and cortical thickness in recurrent major depressive disorder. Sci Rep. 2016;6:21089.

Niu M, Wang Y, Jia Y, Wang J, Zhong S, Lin J, et al. Common and Specific Abnormalities in Cortical Thickness in Patients with Major Depressive and Bipolar Disorders. EBioMedicine. 2017.

Nortje G, Stein DJ, Radua J, Mataix-Cols D, Horn N. Systematic review and voxel-based meta-analysis of diffusion tensor imaging studies in bipolar disorder. J Affect Disord. 2013;150:192-200.

Passos IC, Mwangi B, Cao B, Hamilton JE, Wu MJ, Zhang XY, et al. Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach. J Affect Disord. 2016;193:109-16.

Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. Neuroimage Clin. 2016;10:115-23.

Paulus MP. Pragmatism Instead of Mechanism: A Call for Impactful Biological Psychiatry. JAMA Psychiatry. 2015;72:631-2.

Phillips ML, Swartz HA. A critical appraisal of neuroimaging studies of bipolar disorder: toward a new conceptualization of underlying neural circuitry and a road map for future research. Am J Psychiatry. 2014;171:829-43.

Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. Nat Neurosci. 2014;17:1510-7.

Poline JB, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, et al. Data sharing in neuroimaging research. Front Neuroinform. 2012;6:9.

Rakic P. Confusing cortical columns. Proceedings of the National Academy of Sciences of the

United States of America. 2008;105:12099-100.

Ruggero CJ, Carlson GA, Kotov R, Bromet EJ. Ten-year diagnostic consistency of bipolar disorder in a first-admission sample. Bipolar Disord. 2010;12:21-31.

Ruggero CJ, Kotov R, Carlson GA, Tanenberg-Karant M, Gonzalez DA, Bromet EJ. Diagnostic consistency of major depression with psychosis across 10 years. J Clin Psychiatry. 2011;72:1207-13.

Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry. 2006;163:1905-17.

Sacchet MD, Prasad G, Foland-Ross LC, Thompson PM, Gotlib IH. Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. Front Psychiatry. 2015;6:21.

Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507-17.

Samuel AL. Some studies in machine learning using the game of checkers. IBM Journal of research and development. 1959;3:210-29.

Schmaal L, Hibar DP, Samann PG, Hall GB, Baune BT, Jahanshad N, et al. Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. Mol Psychiatry. 2016.

Schnack HG, Nieuwenhuis M, van Haren NE, Abramovic L, Scheewe TW, Brouwer RM, et al. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. Neuroimage. 2014;84:299-306.

Serpa MH, Ou Y, Schaufelberger MS, Doshi J, Ferreira LK, Machado-Vieira R, et al. Neuroanatomical classification in a population-based sample of psychotic major depression and bipolar I disorder with 1 year of diagnostic stability. Biomed Res Int. 2014;2014:706157.

Shlens J. A tutorial on principal component analysis. arXiv preprint arXiv:14041100. 2014.

Shrager J, Tenenbaum JM. Rapid learning for precision oncology. Nat Rev Clin Oncol. 2014;11:109-18.

Steele JD, Bastin ME, Wardlaw JM, Ebmeier KP. Possible structural abnormality of the brainstem in unipolar depressive illness: a transcranial ultrasound and diffusion tensor magnetic resonance imaging study. J Neurol Neurosurg Psychiatry. 2005;76:1510-5.

Stigler SM. Thomas Bayes's Bayesian Inference. Journal of the Royal Statistical Society Series A (General). 1982;145:250-8.

Stringaris A. Editorial: Neuroimaging in clinical psychiatry--when will the pay off begin? J Child Psychol Psychiatry. 2015;56:1263-5.

Thompson PM, Andreassen OA, Arias-Vasquez A, Bearden CE, Boedhoe PS, Brouwer RM, et al. ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. Neuroimage. 2017;145:389-408.

Tipping ME. Sparse Bayesian learning and the relevance vector machine. Journal of machine

learning research. 2001;1:211-44.

Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Comput Struct Biotechnol J. 2017;15:26-47.

Uher R, Perlis RH, Henigsberg N, Zobel A, Rietschel M, Mors O, et al. Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. Psychol Med. 2012;42:967-80.

Usher J, Leucht S, Falkai P, Scherk H. Correlation between amygdala volume and age in bipolar disorder - a systematic review and meta-analysis of structural MRI studies. Psychiatry Res. 2010;182:1-8.

Wei R, Li C, Fogelson N, Li L. Prediction of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using MRI and Structural Network Features. Front Aging Neurosci. 2016;8:76.

Winkler AM, Kochunov P, Blangero J, Almasy L, Zilles K, Fox PT, et al. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. Neuroimage. 2010;53:1135-46.

Wise T, Cleare AJ, Herane A, Young AH, Arnone D. Diagnostic and therapeutic utility of neuroimaging in depression: an overview. Neuropsychiatr Dis Treat. 2014;10:1509-22.

Wise T, Radua J, Nortje G, Cleare AJ, Young AH, Arnone D. Voxel-Based Meta-Analytical Evidence of Structural Disconnectivity in Major Depression and Bipolar Disorder. Biological psychiatry. 2016a;79:293-302.

Wise T, Radua J, Via E, Cardoner N, Abe O, Adams TM, et al. Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis. Mol Psychiatry. 2016b.

Woo YS, Shim IH, Wang HR, Song HR, Jun TY, Bahk WM. A diagnosis of bipolar spectrum disorder predicts diagnostic conversion from unipolar depression to bipolar disorder: a 5-year retrospective study. J Affect Disord. 2015;174:83-8.

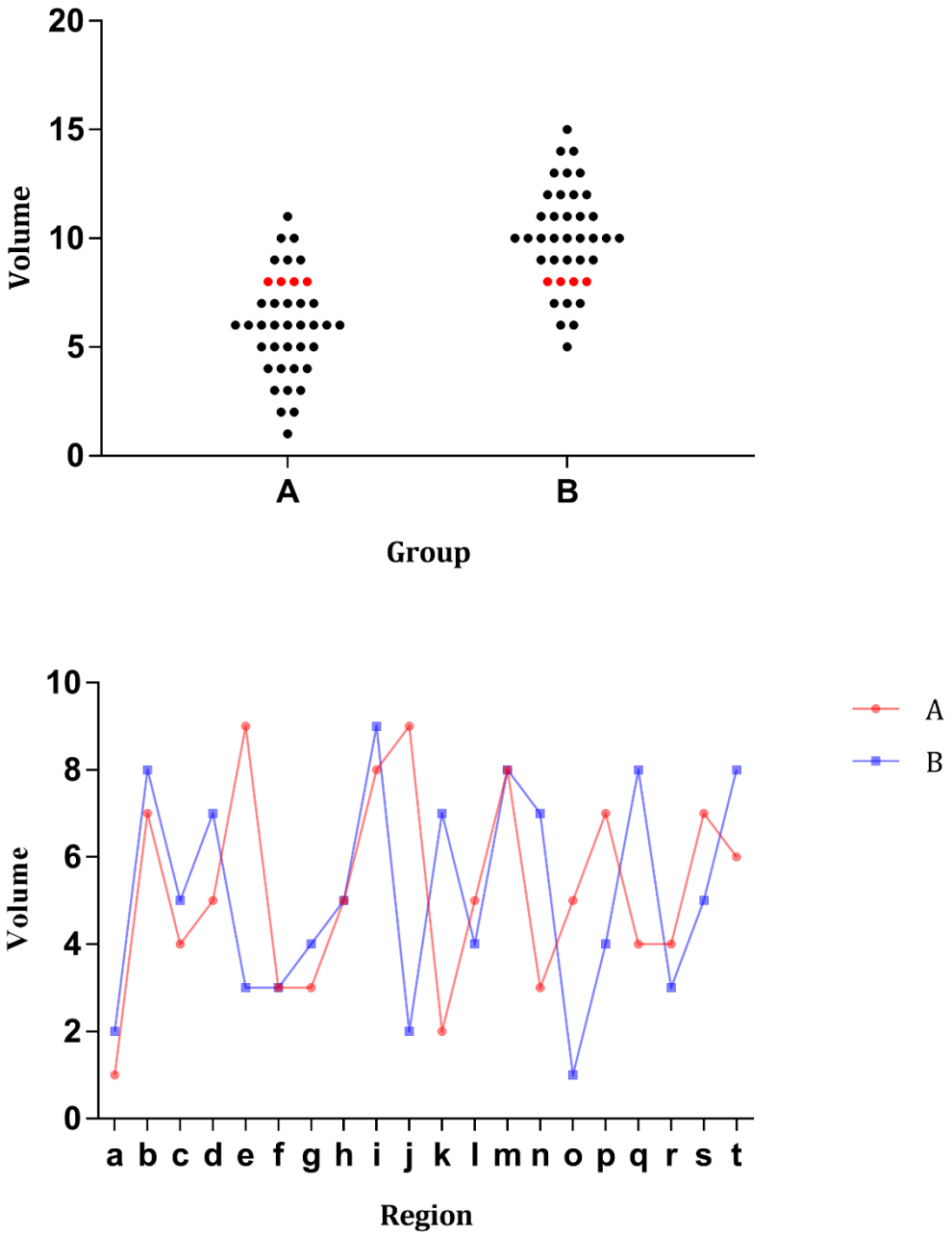
Wray NR, Gottesman, II. Using summary data from the danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. Frontiers in genetics. 2012;3:118.

Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S, et al. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. Neuroimage Clin. 2013;2:735-45.

Zhao L, Wang Y, Jia Y, Zhong S, Sun Y, Zhou Z, et al. Cerebellar microstructural abnormalities in bipolar depression and unipolar depression: A diffusion kurtosis and perfusion imaging study. J Affect Disord. 2016;195:21-31.

Zheng H, Zheng P, Zhao L, Jia J, Tang S, Xu P, et al. Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine. Clin Chim Acta. 2017;464:223-7.

Figure 1. Univariate group-level analysis versus multivariate individual-level analysis

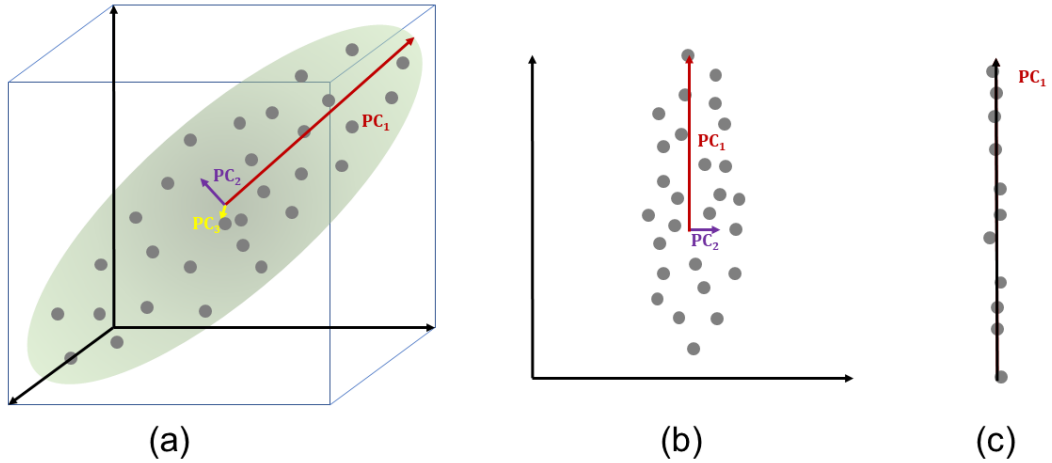


The upper figure represents comparison of a regional structural volume between MDD (A) and healthy controls (B) by univariate group-level analysis. There was a significant statistical difference in the volumes between the two groups ($t = 7.514$, $p < 0.001$). However, individuals who had the same volume could not be classified into group A or B solely based

on volume.

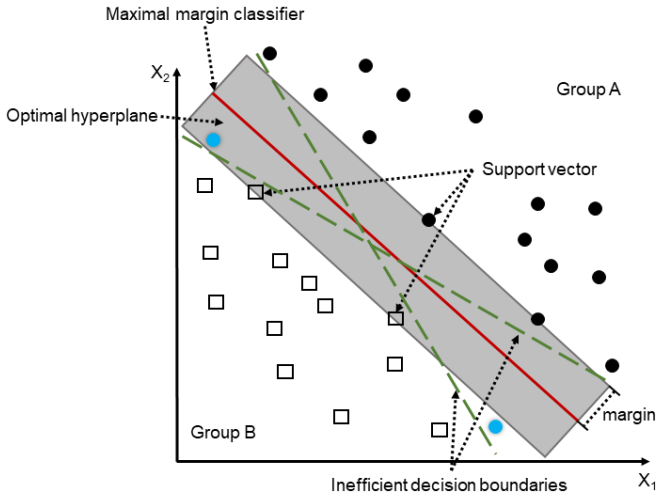
The lower figure shows an example of data from a multivariate individual-level analysis. Each point represents the volume in each brain region (a) to (t). There were three regions in which a patient and healthy individual had the same point, (f), (h), and (m). However, if we have a model to analyze and classify the lines into two groups, then the model can help identify whether a newly visited individual has a mood disorder or not.

Figure 2. Schematic view of the dimensionality reduction in a PCA



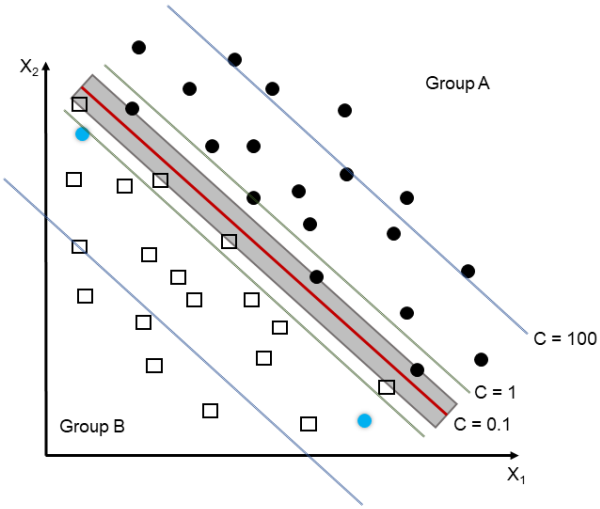
In this figure (a), variables (gray dots) are distributed in a 3-dimensional space. During the process of PCA, three principal components (PC) were generated. The PC_1 explains the largest amount variance among the variables, followed by PC_2, and PC_3. PCs which explain too little of the variance are discarded. (b), PC_3 was discarded due to its minimal explanation of the variance among the variables. Thus, the 3-dimensionality (a) was reduced to 2-dimensionality (b). (b) shows variables explained by PC_1 and PC_2 projected onto the 2-dimensional PCA space. (c) shows 1-dimensionality in which only PC_1 explains the major amount of variance among the variables.

Figure 3. Schematic representation of a support vector machine



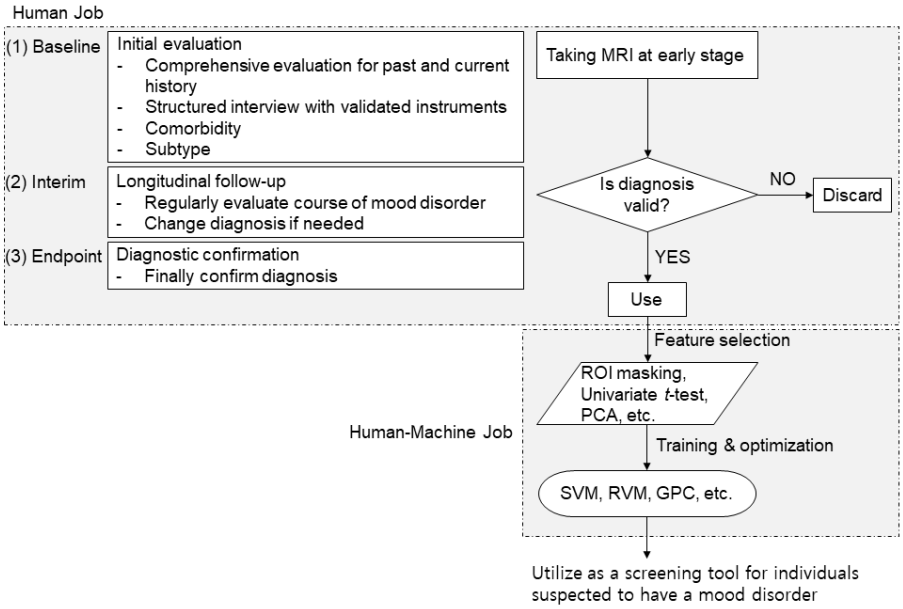
The black circles and rectangles represent each group to be classified. The black circles and rectangles on the boundary of the shadowed zone indicate support vectors. There could be several linear classifiers which divide data into the two groups. Based on the lines, upper and lower data points are classified into the A and B group, respectively. The red line indicates the maximal margin classifier. In case of a 'hard margin', this could be too narrow. For example, when the left-upper blue circle is in the Group A, the classifier should be under the circle like the dashed green line. It is the same in the case of the right-lower blue circle and underlying dashed green line.

Figure 4. Schematic representation of the role of parameter C in a support vector machine



The above figure shows three ranges of decision boundaries with different values for parameter C. As the C values decreases, the margin and error rates decrease and the risk of over-fitting increases. Whereas, as the C value increases, then the margin becomes widened, error rate increases, and the risk of over-fitting decreases.

Figure 5. Summary of the machine learning process.



The essential of machine learning is not the 'machine' but the 'human'. To prevent 'trash-in, trash-out', careful recruitment, evaluation, and selection of participants in the initial stage is substantially important. Given the possibility of changes in the diagnosis in mood disorder, longitudinal follow-up is recommended, although the brain sMRI is taken at the baseline to obtain unbiased data.

Table 1. Summary of key elements in the machine learning process with sMRI

Feature selection methods	To reduce dimensionality, discard unimportant variables and preserve essential features.
PCA	Unsupervised methods. Generating artificial variables based on attribution to the whole feature. Artificial variables with less effect on the data are discarded.
Univariate *t*-tests	Discard non-significant regions based on the results from group-level univariate *t*-tests.
Masking ROI	Select regions to be included or excluded based on knowledge and evidence derived from previous studies.
Classifier	
SVM	The most widely used classifier in machine learning for psychiatric disorders. Adjust balance between bias and variance by a parameter C. Kernel function enables SVM to work on multiple dimensions.
Bayesian model	Uses a probabilistic parameter and output probability ranging from 0 to 1, but not yes/no just like in the SVM. RVM and GPC are examples of the Bayesian model.

PCA: principal component analysis. ROI: region of interest. SVM: support vector machine. RVM: relevant vector machine. GPC: Gaussian Process Classification

Table 2. Clinical characteristics of machine learning studies for mood disorders

Author(Year)	Age	Number (F)	Age at onset	Severity	Number of episodes	Duration of illness	Psychotropic medication	Comorbidity	Longitudinal follow-up	Accuracy
MDD vs HC										
(Costafreda et al., 2009)	43.2 (8.8)	37(28)	NA	HRSD 20.7 (2.2)	NA	NA	Free for a minimum 4 weeks (8 weeks for fluoxetine) at baseline	Exclusion of Axis I (including history of substance abuse within 2 months) and Axis II	None	67.6 (0.027)
(Gong et al., 2011)	39.17 (12.88)	23(13)	NA	HRSD 24.22 (3.76)	NA	30.61 (35.85) m	Drug-naïve	Exclusion of previous psychiatric treatment, lifetime alcohol or drug abuse	None	84.65 (<0.001)
(Liu et al., 2012)	26.71 (7.73)	17 (7)	NA	HRSD 25.58 (6.32)	1st episode	2.59 (1.33)	Drug-naïve	Exclusion of bipolar disorder, neurological illness, or a lifetime history of alcohol or drug use	None	91.2 (NA)
(Mwangi et al., 2012)[3]	46.1 (12.5)	15 (9)	NA	HRSD 23.2 (4.3)	NA	Minimum > 3 months, a first episode diagnosed at 5 years before recruitment	Stable for at least 1month before scanning	Exclusion of any other psychiatric diagnosis such as personality disorder, history of substance misuse	NA	90.3 (< 1 x 10^{-7})
(Serpa et al., 2014)	29.1 (8.34)	19 (15)	NA	NA	1st episode	250.8 (205.7) d	Antipsychotics (n = 15), mood stabilizers	Psychotic MDD. Exclusion of	1 year	59.6 (> 0.05)

							(n = 4), antidepressants (n = 10)	moderate or severe mental retardation		
(Johnston et al., 2015)	51.8 (11.23)	20 (15)	NA	HRSD 16.10 (5.58) MADRS 22.50 (7.97)	Lifetime and/or concurrent chronic episode	>24m	One or more antidepressants (n = 18), additional antipsychotics (n = 3) and lithium (n = 3)	TRD. Exclusion of other primary psychiatric disorder, history of current or previous substance misuse	None	85 (<0.0001)
(Sacchet et al., 2015)	35.6 (8.4)	14 (14)	16.3 (6.8)	HRSD 18.6 (4.2)	NA	18.2 (11.0) y	One or more psychotropic drugs (n = 3)	Various comorbidities (one or more anxiety disorders, n = 7)	None	71.88 (<0.025)
(Jie et al., 2015)	20.1 (2.8)	25 (16)	NA		NA	NA	Medications were unchanged for three weeks prior to scanning.	Exclusion of active substance use disorder, PTSD, OCD, family history of BD	None	62.50 (NA)
Bipolar disorder (vs MDD)										
(Serpa et al., 2014)	27.1 (8.87)	23 (14)	NA	NA	1st episode	184.5 (130.7)	Antipsychotics (n = 10), mood stabilizers (n = 12), antidepressants (n = 1)	Psychotic mania of BID. Exclusion of moderate or severe mental retardation	1 year	54.76 (> 0.05)
(Jie et al., 2015)	21.6 (2.9)	21 (9)	NA	YMRS 1.7 (2.2)	NA	NA	Medications were unchanged for three weeks prior to scanning.	Exclusion of active substance use disorder, PTSD, OCD	None	76.09 (NA)
Bipolar disorder (vs										

schizophrenia) (Schnack et al., 2014)	37.7 (11.0)	66 (42)	86.4	NA	NA	12.8 (9.7) y	Antipsychotics (n = 12), lithium (n = 45)	NA	None	86.4 (NA)

HIGHLIGHT

- Machine learning approach for the neuroimaging data is useful for diagnosing individual patients.
- Many of current machine learning studies did not consider clinical correlates.
- Future machine learning studies should integrate variable clinical and neurobiological factors.