PERSPECTIVE

Clinical application of brain imaging for the diagnosis of mood disorders: the current state of play

JB Savitz[1,2], SL Rauch[3,4] and WC Drevets[1,5]

In response to queries about whether brain imaging technology has reached the point where it is useful for making a clinical diagnosis and for helping to guide treatment selection, the American Psychiatric Association (APA) has recently written a position paper on the Clinical Application of Brain Imaging in Psychiatry. The following perspective piece is based on our contribution to this APA position paper, which specifically emphasized the application of neuroimaging in mood disorders. We present an introductory overview of the challenges faced by researchers in developing valid and reliable biomarkers for psychiatric disorders, followed by a synopsis of the extant neuroimaging findings in mood disorders, and an evidence-based review of the current research on brain imaging biomarkers in adult mood disorders. Although there are a number of promising results, by the standards proposed below, we argue that there are currently no brain imaging biomarkers that are clinically useful for establishing diagnosis or predicting treatment outcome in mood disorders.

OVERVIEW OF APPLICATIONS OF NEUROIMAGING IN PSYCHIATRIC DISORDERS

The application of neuroimaging technology in psychiatric research has revolutionized clinical neuroscience perspectives on the pathophysiology of the major psychiatric disorders. Research using a variety of types of neuroimaging techniques has shown that these conditions are associated with abnormalities of brain function, structure and receptor pharmacology. These data also corroborate the conclusions reached from genetic, endocrine, and clinical pharmacology research involving these disorders to suggest that under the current nosology the major psychiatric disorders likely reflect heterogenous groups of disorders with respect to pathophysiology and etiology.

Despite the invaluable leads that the neuroimaging studies have provided regarding the neurobiological bases for psychiatric disorders, they have yet to impact significantly the diagnosis or treatment of individual patients. In clinical medicine, considerable interest has existed in developing objective, biologically based tests for psychiatric illnesses. From the clinical perspective, such advances could yield important benefits such as predicting treatment response, differentiating between related diagnostic categories, and potentially treating at-risk patients prophylactically to prevent neurotoxicity and clinical deterioration.

Nevertheless, the effect size of neuroimaging and other biological abnormalities identified to date in psychiatric disorders has been relatively small, such that imaging measures do not provide sufficient specificity and sensitivity to accurately classify individual cases with respect to the presence of a psychiatric illness. This review focuses specifically on the *potential* clinical utility of biomarkers assessed using modern neuroimaging technologies, and the *approach* required to validate imaging biomarkers for use as clinical diagnostics.

The quest for biomarkers in psychiatry

Both the clinical practice of psychiatry and the development of novel therapeutics have been hindered by the lack of biomarkers that can serve as accessible, objective indices of the complex biological phenomena that underpin psychiatric illness. The inaccessibility of brain tissue, the lack of knowledge about pathophysiology and the uncertain link between abnormal measurements on any biological test and pathogenesis all have impeded the development of biomarkers for psychiatric disorders. As a result, progress toward improving diagnostic capabilities and defining or predicting treatment outcome in psychiatry has lagged behind other areas of medicine. Thus, it frequently remains difficult to establish whether individual patients suffer from a particular disease, how individual patients can best be treated, and whether experimental treatments are effective in general.

The need for clinical biomarkers has become acute, as their absence particularly has hindered research aimed at developing novel therapeutics. Due at least partly to the lack of well-established pathophysiological targets for new drugs, relatively large numbers of experimental compounds are failing in increasingly expensive late-stage clinical trials. As a result, drug development pipelines are becoming dry, and several companies have discontinued their research and development of pharmaceuticals for psychiatric conditions. The ramifications of these limitations for clinical practice also are significant, as psychiatric nosology and diagnosis largely have remained at a standstill. Since the development of Diagnostic and Statistical Manual of the American Psychiatric Association (DSM)-III, the clinical approach to treatment decisions for individual patients remains empirical ('trial and error'), and many patients are inadequately helped by extant treatments.

[1]Laureate Institute for Brain Research, Tulsa, OK, USA; [2]Department of Medicine, Tulsa School of Community Medicine at the University of Tulsa, Tulsa, OK, USA; [3]McLean Hospital, Belmont, MA, USA; [4]Department of Psychiatry, Harvard Medical School, Boston, MA, USA and [5]Johnson and Johnson Pharmaceuticals Research and Development, Titusville, NJ, USA. Correspondence: Professor W Drevets, Laureate Institute for Brain Research, 6655 South Yale Avenue, Tulsa, OK 74136, USA.
E-mail: wdrevets@laureateinstitute.org

Current application of neuroimaging biomarkers in psychiatric diagnosis

For over two decades, imaging has maintained a well-established but narrow place in the diagnostic evaluation of patients with psychiatric disease, largely because of the usefulness of neuromorphological magnetic resonance imaging (MRI) in detecting and characterizing structural brain abnormalities such as lesions and atrophy. Thus, the role of imaging in patients with psychopathology historically has been limited to one of exclusion of potentially etiological medical conditions: namely to rule out neoplasm, hematoma, hydrocephalus or other potentially surgically treatable causes of psychiatric symptoms, or to detect the presence of cerebrovascular disease or gross atrophy. Although clinically important, these conditions appear to have a role in the pathogenesis of psychiatric symptoms in only a small proportion of cases presenting for the evaluation of mood, anxiety or psychotic disorders.

Increasingly, a major quest of researchers has been to identify neuroimaging results that offer diagnostic capabilities for particular psychiatric diseases as well as for their relevant differential diagnoses. Currently, neuroimaging is not recommended within either the US or the European practice guidelines for positively defining diagnosis of any primary psychiatric disorder. Nevertheless, advances in *research* applications of neuroimaging technology have provided leads that may foreshadow future *clinical* applications of imaging biomarkers for establishing diagnosis and predicting illness course or treatment outcome. The ensuing review discusses issues that have been addressed within other areas of clinical medicine to establish the validity and reliability of imaging diagnostics, with the aim of providing principles to guide the evaluation of neuroimaging applications in clinical psychiatry.

BIOMARKER DEFINITION, VALIDATION AND QUALIFICATION

The National Institutes of Health has defined a biomarker (that is, biological marker) as: 'A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes or pharmacologic responses to a therapeutic intervention.'[1] A biomarker thus can define a physiological, pathological or anatomical characteristic or measurement that putatively relates to some aspect of either normal or abnormal biological function. Biomarkers thus may assess many different types of biological characteristics or parameters, including receptor expression patterns, radiographic or other imaging-based measures, or electrophysiologic parameters.

Furthermore, the term 'biomarker' connotes different meanings in different contexts, based upon the intended application of the information a biomarker provides. Within clinical medicine, biomarkers include measures that suggest the etiology of, susceptibility to, activity levels of, or progress of a disease. In addition, alterations in patient-associated biomarkers related to an intervention may be used to predict the likelihood of experiencing a robust clinical outcome or an adverse reaction to a treatment. Finally, in drug development a biomarker can be any measure of drug action that is proximal to its clinical effect, including biomarkers that correlate with drug response or quantify the extent to which a drug occupies specific receptors in a target tissue.

Notably, the US Food and Drug Administration (FDA) and the European Medicines Agency recently have jointly developed guidance that addresses multiple types of biomarkers that can be applied to drug development, including prognostic, predictive, pharmacodynamic and surrogate biomarkers. A *prognostic biomarker* is a baseline patient or disease characteristic that categorizes patients by degree of risk for disease occurrence or progression. A *predictive* biomarker is a baseline characteristic that categorizes patients by their likelihood for response to a particular treatment. A *pharmacodynamic* biomarker is a dynamic assessment that shows that a biological response has occurred in a patient after having received a therapeutic intervention. A *surrogate* end point is defined as a biomarker intended to substitute for a clinical efficacy end point. Conceivably, each of these biomarker types holds the potential to be clinically useful in psychiatric research or practice. Nevertheless, in its guidance the FDA identified the most valuable role for biomarkers as their use in clinical *diagnostics*.

In considering the development of neuroimaging biomarkers as clinical diagnostics, the FDA guidance on biomarkers for drug development merits comment. Generally, the requirements of biomarkers for quantification of drug effects in research and development, which depend upon population means with variance estimates, converge with the requirements of diagnostics in clinical practice, which are assessed on a per-patient basis. The common element in both is longitudinal quantification; both analyses require baseline and follow-up effects of treatments. For example, clinical evidence from the National Oncologic positron emission tomography (PET) Registry motivated the expanded coverage by Medicare for fluorodeoxyglucose-PET/CT (computed tomography) in the detection and staging of cancer and in the monitoring of cancer treatment response. Thus as diagnostics, biomarkers are of interest to health-care providers and consumers for parallel applications, since earlier detection of disease facilitates earlier intervention, which, when followed by effective, individualized treatment, can improve patient outcomes.

With respect to establishing the utility of a biomarker, it is useful to distinguish between the terms 'validation' and 'qualification'. *Validation* generally refers to the determination of the performance characteristics of a measurement—for example, the measurement's reliability, sensitivity and specificity—in measuring a particular biological construct. The validation process is particularly relevant for securing regulatory approval to market techniques for commercial use as clinical diagnostics, as described in the subsequent section.

The term *qualification* refers to the establishment of the credibility of a biomarker in its application to questions specifically relevant to drug development. In drug development, the ultimate use of a biomarker is as a *surrogate* end point, which requires that the biomarker has been *qualified* to substitute for a clinical standard of truth (that is, the biomarker reasonably predicts the clinical outcome and therefore can serve as a surrogate). After a biomarker is 'qualified' by the FDA (or other regulatory agency), industry can use the markers in a similar context in multiple drug trials, drug classes or clinical disorders, without having to repeatedly seek the agency's approval ('Qualification Process for Drug Development Tools'; http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf).

The FDA *qualification* process for biomarkers also encompasses guidance on drug-development tools, including radiographic or other imaging-based measurements. Qualification of a drug-development tool is based on a conclusion that within the stated context of use, the results of assessment with the tool can be relied upon to have a specific interpretation and application under regulatory review. The FDA guidance indicates 'While a biomarker cannot become qualified without a reliable means to measure the biomarker, FDA clearance of a measurement device does not imply that the biomarker has been demonstrated to have a qualified use in drug development and evaluation.' Instead the qualification process is limited to specific patient populations and a specific therapeutic intervention. In addition to the biomarker assay *validation* data, clinical data are required to support the biomarker *qualification*. A corollary of this regulatory principle is that the FDA qualification of a drug-development tool for one application does not extend to its use in other applications.

530

EVALUATING THE VALIDITY OF DIAGNOSTIC BIOMARKERS IN CLINICAL MEDICINE

The validity of a diagnostic biomarker for any medical disorder generally is established via evaluation of its sensitivity, specificity, prior probability, positive predictive value and negative predictive value.[2] *Sensitivity* refers to the capacity of a biomarker to identify a substantial percentage of patients with the disease-of-interest (expressed as: true positive cases divided by (true positive cases plus false negative cases) × 100). Thus, a sensitivity of 100% corresponds to a marker that identifies 100% of patients with the target condition. *Specificity* refers to the capacity of a test to distinguish the target condition from normative conditions (for example, aging) and other pathological conditions (expressed as: true negatives divided by (true negative cases plus false positive cases) × 100). A test with 100% specificity would be capable of differentiating the target condition from other conditions in every case. *Prior probability* is defined as the frequency of occurrence of a disease in a particular population (true positives plus false negatives divided by the total population). A perfect biomarker would detect only true positives and no false negatives and thus would reflect accurately the prevalence of the disease in the population. *Positive predictive value* is the percentage of people who have a positive test who can be shown by a definitive examination (for example, subsequent autopsy or biopsy) to have the disease (true positives divided by (true positives plus false positives)). A positive predictive value of 100% indicates that all patients with a positive test actually have the disease. For a biomarker to be considered useful clinically, it generally is expected to show a positive predictive value of ∼ 80% or more.[3] *Negative predictive value* represents the percentage of people with a negative test that subsequently proves not to have the disease on definitive examination (true negatives divided by (true negatives plus false negatives)). A negative predictive value of 100% indicates that the test completely rules out the possibility that the individual has the disease, at least at the time the individual is tested. A reliable marker with a high negative predictive value is extremely useful in clinical medicine, although a test with low negative predictive value can in some cases still be useful if it also has high positive predictive value.

In the development of medical tests, the threshold for distinguishing abnormal from normal alters the sensitivity and specificity in opposite ways. Thus, if the threshold is set further from the distribution of normative values then the test becomes less sensitive for detecting true positives, but more specific for rejecting true negatives. The convention in establishing diagnostic tests for medical conditions has been to select an intermediate choice that minimizes the total error from both false positives and false negatives.

In the case of Alzheimer's Disease (AD), the Consensus Report of the Working Group on Molecular and Biochemical Markers of Alzheimer's Disease,[3] for example, recommended that in order to qualify as a biomarker the measurement in question should detect a fundamental feature of neuropathology and be validated in neuropathologically confirmed cases, and should have a sensitivity of >80% for detecting AD and a specificity of >80% for distinguishing other dementias.[3] The validation of diagnostic biomarkers for AD has been facilitated by the capability for confirming the diagnosis *post mortem*. Thus, the current clinical criteria for returning a diagnosis of 'probable AD' provide a sensitivity of about 85% when compared with autopsy-confirmed cases. In order for a diagnostic *biomarker* to be clinically *useful*, therefore, its sensitivity must exceed this value when correlated with neuropathology (otherwise there is no benefit to performing the test). For example, the validation of a diagnostic neuroimaging marker for β-amyloid pathology in AD, [F-18] florbetapir, is being evaluated partly on the basis of correlating florbetapir-PET data

acquired *antemortem* with evidence of β-amyloid in the same subjects *post mortem*. The results rated as positive or negative for β-amyloid agreed in 96% of 29 individuals assessed in the primary analysis cohort. In a secondary analysis, non-autopsy cohort, florbetapir-PET images were rated as amyloid negative in 100% of 74 younger individuals who were cognitively normal,[4] suggesting that negative results on this test hold high negative predictive value.

Nevertheless, the outcome of the FDA evaluation of [F-18] florbetapir-PET for commercial use as a clinical diagnostic tool illustrates another central principle in the validation of an imaging diagnostic biomarker, namely that the reliability of ratings across radiologists must be relatively high. In January 2011, the Peripheral and Central Nervous System Drugs Advisory Committee of the FDA recommended *against* approval of the new drug application for [F-18]florbetapir injection, based largely on concerns about the variability of ratings across readers. The Advisory Committee chair said during an interview after the meeting, 'We would like to see some structured training and evidence of consistency among readers' (http://www.medscape.com/viewarticle/739297). In the pivotal trial described in the previous paragraph, Clark et al.[4] used the median of three readers' visual ratings on a five-point scale to assign the extent to which the PET scan was positive for amyloid protein binding. Since inspection of the data from individual readers ultimately raised questions about inter-rater reliability, the FDA response focused primarily on the need to establish a reader-training program for market implementation that would serve to ensure reader accuracy and consistency of interpretation of existing [F-18]florbetapir scans. The FDA subsequently approved Amyvid (Florbetapir F-18 Injection) for use in patients being evaluated for AD and other causes of cognitive decline in April of 2012. A key aspect of securing approval is that the company who will market and distribute this PET radiopharmaceutical instituted a training program for radiologists who will read the scans to ensure inter-rater reliability.

The need to ensure that readers consistently can detect clear positive or negative results extends to the clinical application of any imaging procedure for which the results depend on the *subjective* interpretation of a reader. For biological assays that can be *objectively* quantified, the accuracy often is characterized by comparing the assay results obtained for a known standard (for example, a test sample with known concentration for the target compound) and the reliability or reproducibility is statistically expressed with respect to the variability in the quantitative results obtained after performing repeated testing on the same sample. In contrast, many types of clinical imaging assessments depend upon subjective interpretation, such as a radiologist's reading of a radiographic or nuclear medicine (for example, PET and single-photon emission computed tomography) image on the basis of gross visual inspection of the image. In this case, the variability of such interpretations is evaluated by characterizing the reliability and variability of the results obtained within and across raters.

Thus, intra-rater reliability can be established by assessing the extent to which readings performed *under blind conditions* by the same reader on the *same image* on different days are in agreement, and/or the extent to which the same radiologist renders the same results when comparing images obtained from the same patient on different days. Similarly, inter-rater reliability is assessed by having multiple radiologists read the same set of images while blind to the evaluations returned by the other readers. These intra-rater and inter-rater reliability assessments thus evaluate, respectively, the intra-individual variability (reflecting the failure of a reader to be consistent with himself or herself) and the inter-individual variability of interpretations (reflecting inconsistency of interpretation among different readers).

CHALLENGES IN ESTABLISHING THE VALIDITY OF DIAGNOSTIC BIOMARKERS IN PSYCHIATRY

An important challenge in the application of neuroimaging to psychiatric diagnosis is that the clinical utility of such tests depends partly upon their ability to distinguish multiple conditions from one other. In general, both the intra-individual and inter-individual variability of interpretation increase in proportion to the number of diagnostic categories that are considered clinically relevant. In other words, the fewer the categories into which readers are assigning results, the greater the degree of agreement between readers. This tendency was illustrated historically by the results of a landmark study that evaluated the variability in interpreting chest X-ray films during lung cancer screening.[5] The study radiologists showed 65.1% agreement when they were required to place the film results into one of five categories (suspected neoplasm, other significant pulmonary abnormality, cardiovascular abnormality, non-significant abnormality and negative), compared with 89.4% agreement if they were instead required to place the results into only two categories (positive or negative for significant pulmonary abnormality). Presumably, a diagnostic biomarker assessment aimed at informing the differential diagnosis of psychiatric disorders would need to address more than two categories, however, increasing the variability of image interpretations across readers.

In psychiatry, the need to differentiate various conditions from each other depends partly on the clinical imperative to return distinct treatment recommendations for different disorders. It might be argued, for example, that for a neuroimaging procedure to add clinical value in the evaluation of an adult patient with impaired attention, the differential diagnosis relevant to the treating physician includes major depressive disorder (MDD), bipolar disorder (BD), attention deficit disorder and anxiety disorders, at a minimum,[6] since the standard of care differs between these categories. Thus, the variability across raters will be relatively higher (that is, lower inter-rater reliability) for a diagnostic imaging study that must differentiate among several psychiatric disorders that share symptomatology but require distinct treatment approaches as compared with the case such as that described above for [F-18]florbetapir-PET, which hinges only on two categories (β-amyloid positive versus negative).

Furthermore, the determinations of positive and negative predictive value are limited by the absence of an established objective standard for establishing diagnosis in psychiatric disease (for example, analogous to the neuropathologically verified diagnosis of AD). Thus, the absence of certain knowledge about the pathophysiology of psychiatric disorders will hinder the development and validation of *diagnostic* biomarkers. Greater optimism has been associated with establishing *predictive* biomarkers of treatment response, *pharmacodynamic* biomarkers of the effect of pharmacological probes, and *surrogate* biomarkers of treatment outcome based on translational studies that ultimately can facilitate the discovery of pathophysiology.

Nevertheless, it might be argued that the Consensus Report of the Working Group on Molecular and Biochemical Markers of Alzheimer's Disease[3] reviewed above offers a template for developing diagnostic biomarkers of psychiatric disease. Of course, the fundamental recommendation that 'in order to qualify as a biomarker the measurement in question should detect a fundamental feature of neuropathology and be validated in neuropathologically confirmed cases' cannot be applied directly to psychiatric disorders. Thus, the psychiatric imaging field is moving forward by establishing gold-standard diagnoses using criteria-based conventions.[6] If this approach for establishing the 'actual' diagnosis is accepted, then the remainder of this Consensus Report can be meaningfully adapted to biomarker validation in psychiatric disorders. This approach would argue that

a diagnostic biomarker should have a sensitivity of >80% for detecting a particular psychiatric disorder and a specificity of >80% for distinguishing this disorder from other psychiatric or medical disorders. The biomarker ideally also should be reliable, reproducible, non-invasive, simple to perform and inexpensive. Finally, the validating data used to establish a biomarker require confirmation by at least two independent sets of qualified investigators with the results published in peer-reviewed journals.

According to this standard, the psychiatric imaging literature currently does not support the application of a diagnostic neuroimaging biomarker to positively establish the presence of any primary psychiatric disorder. Although assessments of intra-rater and inter-rater reliabilities commonly are reported for quantitative neuroimaging measures, these have been limited to establishing *measurement* reliability (for example, of cerebral volumes or neuroreceptor binding potential (BP)), but not to the reliability of diagnostic interpretation. Thus, the peer-reviewed scientific literature does not yet contain an example of a diagnostic imaging biomarker with regard to a psychiatric disorder or treatment for which relatively high intra- and inter-rater reliabilities have been reported in two independent studies or laboratories. Similarly, there is not yet a case in the literature where neuroimaging measures obtained from the same region(s)-of-interest has shown both a sensitivity of >80% for detecting a particular psychiatric disorder and a specificity of >80% for distinguishing this disorder from healthy controls or other relevant psychiatric disorders. Nevertheless, the ensuing sections review progress toward developing such biomarkers using state-of-the-art neuroimaging technologies. Notably, this literature contains several examples of *individual* studies for which sensitivity and specificity approach or exceed 80%, and it is conceivable that some of these findings ultimately may be replicated in independent studies.

SECTION II: PROGRESS TOWARD A DIAGNOSTIC IMAGING BIOMARKER OF DEPRESSION

### A synopsis of neuroimaging abnormalities associated with mood disorders

The neuroimaging literature has recently been extensively reviewed by us and others elsewhere (for example, ref.[7–13]) and here we highlight only the major themes that characterize this large corpus of data.

There is an emerging consensus that depression is characterized by a fundamental mood-congruent processing bias; that is, a greater sensitivity to punishment and an impaired hedonic capacity. In the context of functional MRI (fMRI) studies, this cognitive bias manifests itself in two principal forms. First, relative to healthy individuals, some depressed patients display a greater hemodynamic response in the amygdala to negatively valenced emotional stimuli such as sad faces and/or a reduced hemodynamic response in the amygdala to positive stimuli such as happy faces[14–18] (Figure 1). Second, some depressed patients show a blunted hemodynamic response in the ventral striatum and orbitofrontal cortex to reward stimuli that may be correlated with anhedonia, a core symptom of depression[19–25] (Figure 2).

The genetic variants that increase the risk for developing mood disorders are incompletely penetrant. Stated differently, individuals with genetic risk factors (for example first-degree relatives) for affective disorders do not necessarily become ill. They may, however, share neurobiological traits with affected patients. This type of disease biomarker has been termed as an endophenotype.[26] For example, compared with controls, healthy adolescents with a parent with MDD showed a greater hemodynamic blood oxygen level-dependent (BOLD) response to fearful faces in the amygdala and nucleus accumbens, and a reduced hemodynamic response to happy faces in nucleus accumbens.[27] Similarly, compared with controls, both adolescents with BD and
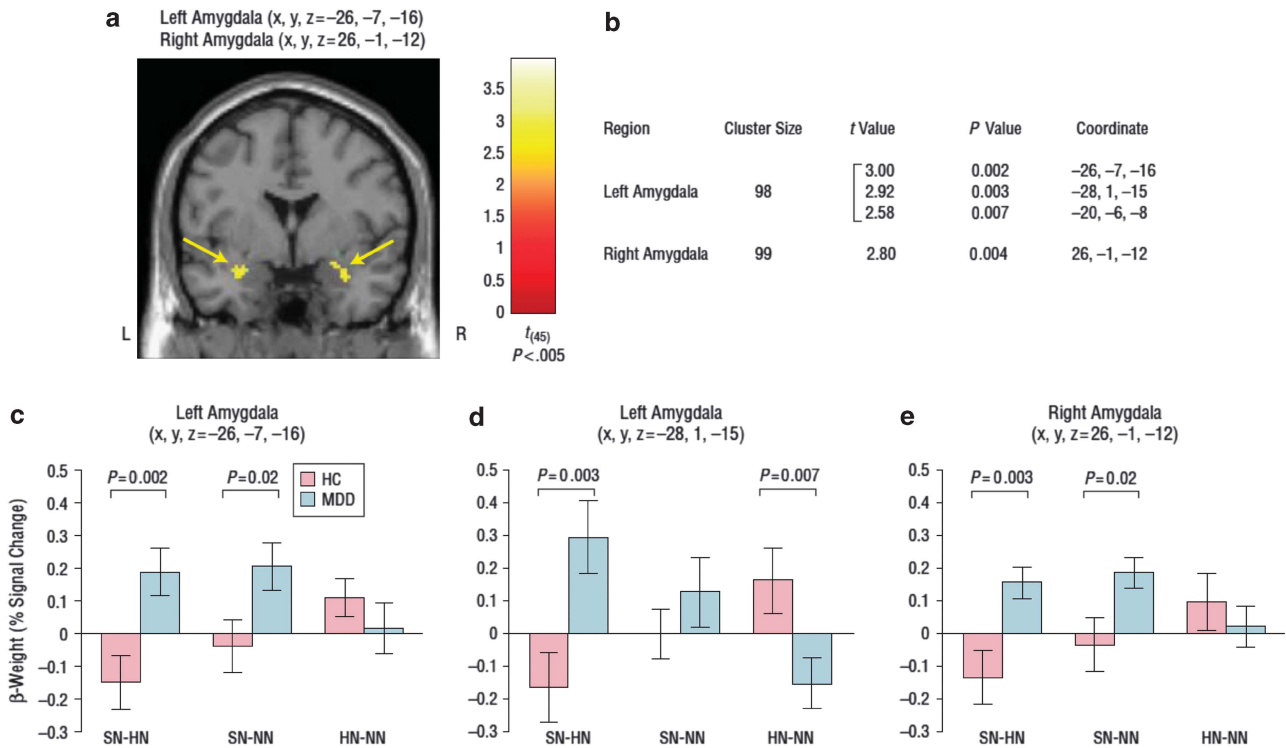
Figure 1. (a) Statistical parametric mapping images consisting of voxel-wise values of the *t*-statistic in the bilateral amygdala indicate differences in the hemodynamic response to masked sad versus masked happy faces (SN-HN) between currently depressed people with major depressive disorder (dMDD) and healthy controls (HCs), shown on a coronal slice located 1 mm posterior to the anterior commissure. (b) Coordinates of peak voxel *t*-value signifying the difference in the amygdala response to SN-HN for dMDD participants versus HCs that correspond to the stereotaxic array of Talairach and Tournoux as the distance in millimeters from the origin (anterior commissure), with positive *x*-value indicating right, positive *y*-value indicating anterior and positive *z*-value indicating dorsal. Cluster size indicates contiguous voxels (*P* < 0.05). Contrast β-weights are shown for specified contrasts in dMDD versus HCs for loci identified in the left (**c**, **d**) and right amygdala (**e**) (reproduced with permission from Victor *et al.*[15]).
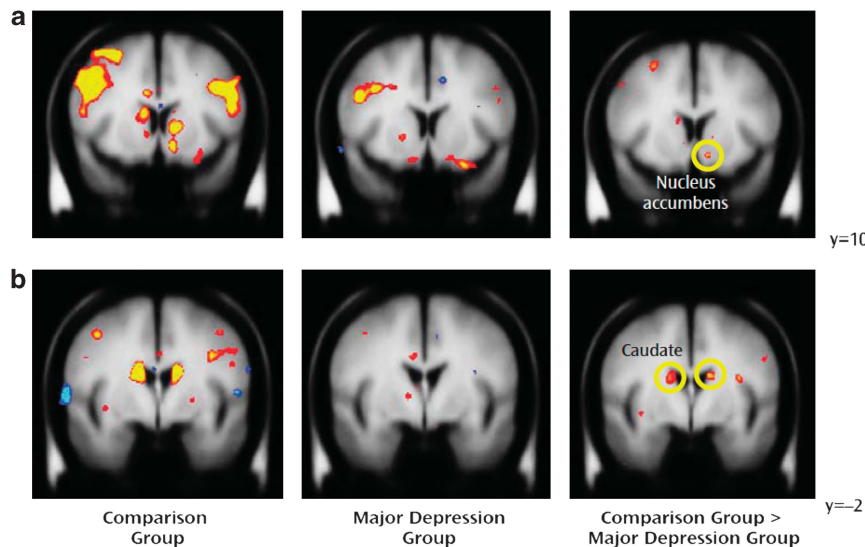
Figure 2. Coronal slices showing consummatory reward activity (monetary gains) in basal ganglia regions are displayed for both comparison subjects and participants with major depression. Relative to the comparison group, the major depression group showed significantly reduced activation in response to gain feedback in the left nucleus accumbens (**a**) and the caudate bilaterally (**b**). All contrasts are thresholded at *P* < 0.005. Left hemisphere is displayed on the viewer's right (reproduced with permission from Pizzagalli *et al.*[25]).

unaffected adolescents with a family-history of BD showed an elevated hemodynamic response in the amygdala when presented with fearful faces.[28] The neurophysiological correlates of response to reward may also serve as an endophenotype of mood disorders. Healthy adolescent girls with a mother with MDD

showed lower activation in the ventral striatum to the anticipation and receipt of a monetary reward,[29] while healthy individuals with a depressed parent displayed an attenuated BOLD response in the orbitofrontal cortex and anterior cingulate cortex (ACC) to a primary reward (the taste of chocolate).[30]
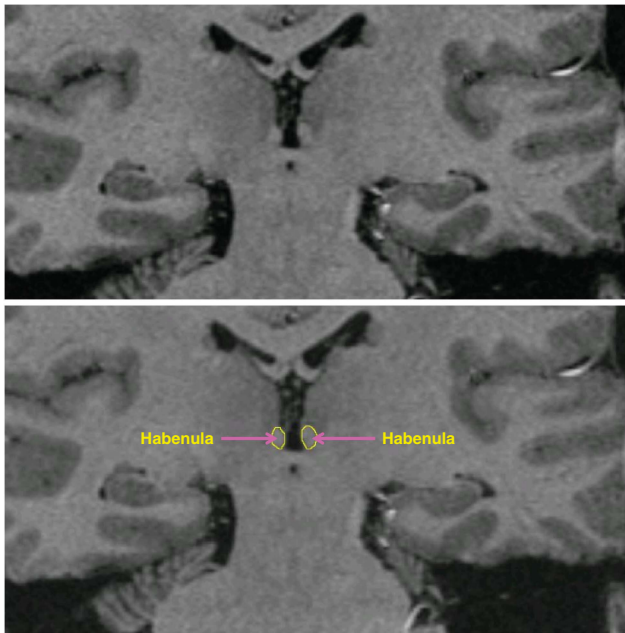
Figure 3. Coronal magnetic resonance imaging (MRI) sections showing the habenula and the local anatomical landmarks that enabled its segmentation. The upper and lower panels show the identical image. The tracing of the habenula is shown in yellow in the lower panel. The small size of the habenula ($\sim 30\,mm^3$) poses significant challenges for the accurate measurement of its volume and functional activity (adapted from Savitz et al.[41,111]).

Neuromorphometric MRI studies of patients with MDD are indicative of reductions in the gray matter (GM) volume of both cortical regions and subcortical structures, especially the subgenual/pregenual ACC,[31–33] the orbitofrontal cortex,[32,34] the hippocampus[34–36] and the striatum.[32,34–36] Many of these findings apply also to patients with BD but these data are less straightforward to interpret because of the neurotrophic effects of mood stabilizers such as lithium which normalize or increase GM volume[37–41] (Figure 3).

Reduced GM volume may constitute an endophenotype for depression. Never-ill adolescent girls with a mother with MDD had significantly lower hippocampal GM density compared with controls[42] while reduced hippocampal, medial prefrontal cortex (PFC) and ACC volume were found in the healthy relatives of patients with MDD.[43] In another study, healthy boys with sub-clinical depressive symptoms were found to have smaller rostral ACC volumes than healthy boys with no depressive symptoms.[44] An objective method for selecting potential endophenotypes from a large set of behavioral, cognitive and morphometric imaging markers identified GM volume changes in several regions including the hypothalamus, hippocampus and pallidum as among the most promising neuroimaging markers for genetic susceptibility to recurrent MDD.[45] Similarly, increased risk of BD was previously associated with GM reductions in the right ACC and ventral striatum.[46]

The reduction in GM volume is hypothesized to result from a loss of neuropil[47] and this deficit is associated with loss of glial cells;[48] each of these findings are hypothesized to arise secondarily to glutamate-induced excitotoxicity.[49] This excitotoxicity hypothesis is partially consistent with the proton magnetic resonance spectroscopy literature, which is indicative of a decrease in Glx (glutamine and glutamate) in the medial PFC and dorsolateral PFC in depressed patients with both MDD and BD.[10,50,51] The decrease in Glx in these regions putatively reflects a decrease in the intracellular component of glutamate and glutamine, a finding that appears consistent with *postmortem*

neuropathological evidence that glial cell counts and density are reduced in depression (reviewed in Ongur et al.[48]), as well as with additional *in vivo* magnetic resonance spectroscopy data showing that the ratio of glutamine to glutamate is reduced in depression.[10] Finally, studies showing that magnetic resonance spectroscopy measures of GABA are abnormally decreased in MDD suggest a decrease in GABAergic signaling[51–53] (Figure 4). The putative depression-associated increase in glutamatergic signaling potentially is consistent with fMRI and PET studies conducted while patients are in the resting state. These data indicate that patients with MDD and BD display elevated glucose metabolism and/or BOLD hemodynamic signal in the region around the genu of the corpus callosum, that is, the perigenual ACC[7,31,54–57] (Figure 5). Interestingly, in subjects with MDD, elevated perigenual ACC activity at rest has also been shown to predictive of a positive response to treatment with antidepressant medications and transcranial magnetic stimulation (reviewed in Pizzagalli et al.[58]).

Extant data indicate that patients with mood disorders show evidence of white matter (WM) pathology as well as reductions in GM volume. WM hyperintensities (WMH) have been observed in a proportion of depressed patients using T2-weighted MRI. In the case of MDD, the prevalence of WMH is elevated significantly in elderly populations who also show a late age-at depression onset.[7] The histopathological correlates and clinical risk factors for WMH in late-onset depression suggest that these MRI-based findings signify cerebrovascular disease when observed within this clinical context, leading to the concept of vascular depression,[59,60] a condition characterized by microvascular disease and/or multiple subcortical infarcts of an ischemic origin. In contrast, there are a number of reports of WMH in both adult and pediatric patients with BD[61–63] for which the etiology remains unknown. Diffusion tensor imaging studies are suggestive of reduced integrity of the WM fibers (particularly within the cingulum and uncinate fasciculi) connecting the PFC with subcortical structures in adult and pediatric patients with BD[64–69] and MDD[70–72] (Figure 6). The reduced depression-associated structural integrity of WM tracts suggested by the diffusion tensor imaging data is paralleled by fMRI studies, which are indicative of reduced functional connectivity between the medial PFC, the dorsolateral PFC and the amygdala when patients are exposed to negatively valenced stimuli.[7,18,73,74]

While there does not appear to be an increase in the number of WMH in the unaffected relatives of patients with BD,[75,76] a reduction in the integrity of WM fibers measured with diffusion tensor imaging has been proposed to be a marker of genetic risk for BD. Children with a first-degree relative with BD demonstrated lower fractional anisotropy (FA), a measure of WM fiber integrity in the superior longitudinal fasciculi[77] and in the corpus callosum and/or inferior fronto-occipital fasciculus of the right temporal lobe[78,79] compared with healthy control children. Further, a study of healthy adults with BD relatives found a generalized reduction in FA throughout the brain.[80] In the case of healthy individuals with a family history of MDD, reduced FA was reported in the cingulum, bilaterally,[81] and the left cingulum, splenium, superior longitudinal fasciculi, uncinate and inferior fronto-occipital fasciculi.[82]

PET technology has enabled researchers to study neuroreceptor function *in vivo* by allowing for the measurement of the BP, which may be heuristically described as the product of the density and affinity of the receptor or protein of interest. Multiple neuroreceptor abnormalities have been reported in mood disorders.[12] Two of the most replicated findings are a reduction in the *postsynaptic* serotonin 1A (5-HT$_{1A}$) receptor BP in the mesiotemporal cortex of patients with MDD[83–87] and BD,[86,88] and an increase in serotonin transporter BP in regions such as the ACC, thalamus and insula in currently depressed patients with MDD[89] and BD,[90] yet not all studies agree with these findings.[91,92]
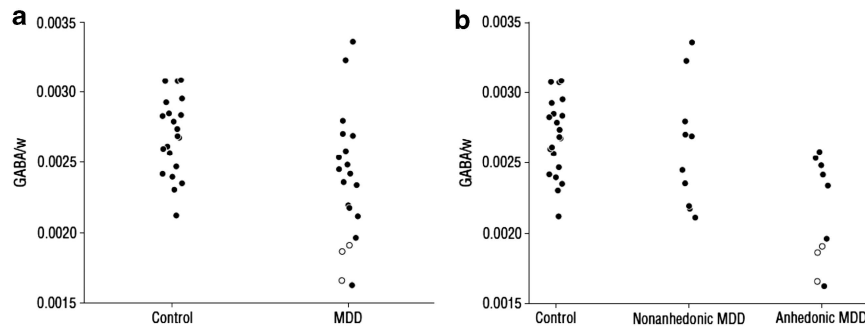
Figure 4. Scatterplots of raw data of anterior cingulate cortex gamma-aminobutyric acid (GABA) relative to unsuppressed voxel tissue water concentrations (GABA/w) in healthy controls and adolescents with major depressive disorder (MDD) (**a**) and healthy controls, non-anhedonic adolescents with MDD, and anhedonic adolescents with MDD (**b**). Open circles represent subjects with melancholic MDD. Note the overlap in the statistical distributions between the mood disorder patients and the healthy controls which is common to all current imaging modalities, and poses challenges for the development of diagnostic tests for mood disorders (reproduced with permission from Gabbay et al.[53]).
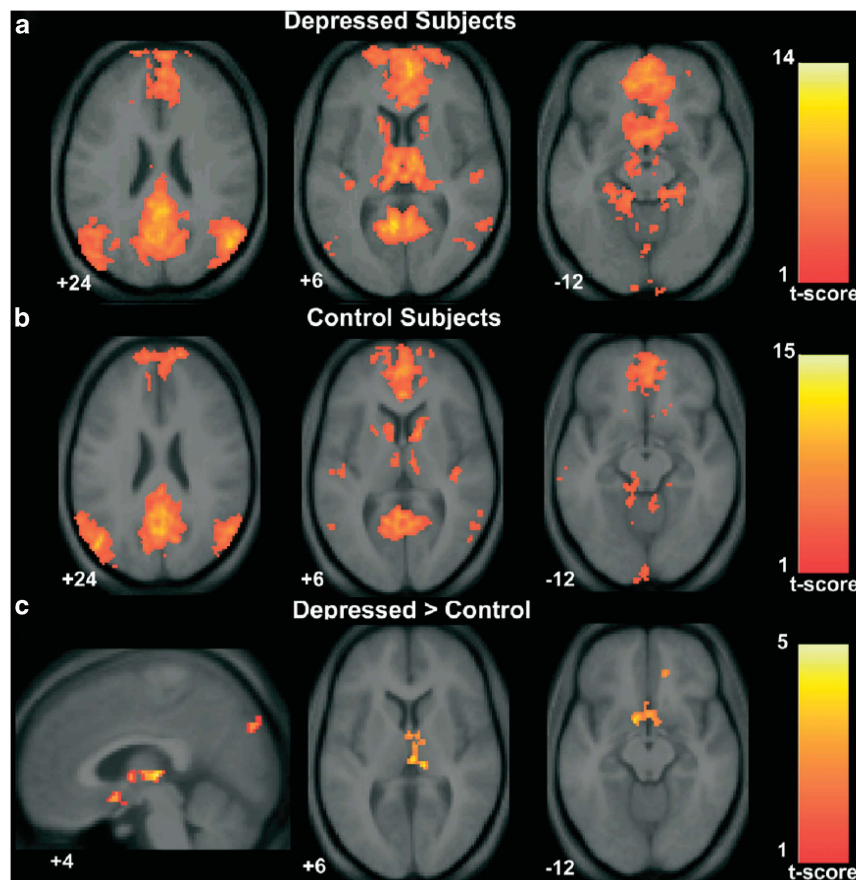

Figure 5. Increased default-mode network functional connectivity in subjects with major depression. Axial images of group default-mode functional connectivity in depressed subjects (**a**) and in healthy controls (**b**). The contrast map in (**c**) demonstrates clusters in the subgenual cingulate, thalamus and precuneus where resting-state functional connectivity was greater in depressed subjects versus controls. The t-score bars are shown at right. Note that while the color scale range begins at 1, the minimum t-values for the analyses were 3.42 for the depressed group map (**a**), 3.58 for the control group map (**b**) and 2.41 for the depressed versus control contrast map (**c**). Numbers at the bottom left of the images refer to the z-coordinates (and for the sagittal image the x-coordinates) in the standard space of the Montreal Neurological Institute (MNI) template. The left side of the image corresponds to the left side of the brain (reproduced with permission from Greicius et al.[57]).

Obstacles to the diagnosis of mood disorders and the development of biomarkers of response to treatment using neuroimaging

Although the broad pattern of neuroimaging abnormalities characteristic of *groups* of patients with mood disorders has been fairly well established, translating these findings into diagnostic tests for the *individual* patient has proven difficult. In

general, the conventional path to validate a diagnostic test is first to generate a potential discriminant function from a patient cohort, and then to test this discriminant function in an *independent* cohort. Currently, to our knowledge, in the case of mood disorders, no such tests have been validated through replication in independent cohorts subject to peer-review.
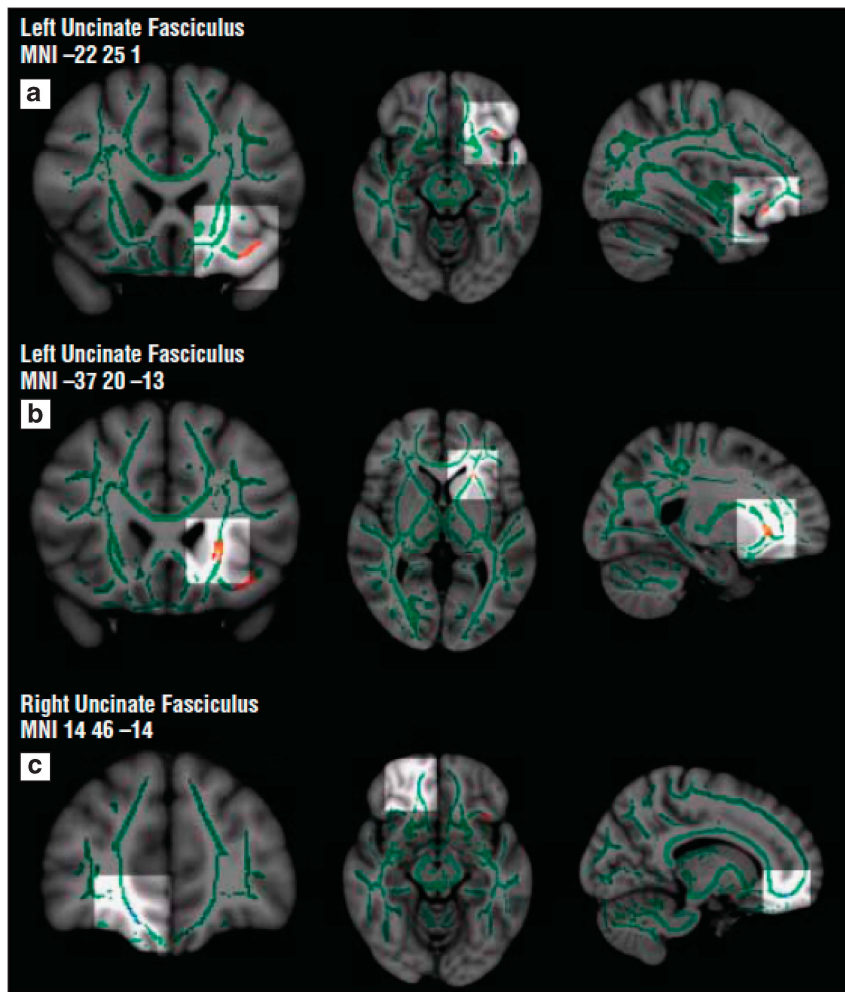
Figure 6. Fractional anisotropy (FA) maps showing (from left to right) coronal, axial and sagittal views. Colored voxels represent regions in which FA differs significantly in subjects with bipolar disorder (BD) versus control subjects. Red-yellow indicates greater FA in subjects with BD versus controls; light blue, decreased FA in subjects with BD versus controls ($t > 3.0$ and $P < 0.05$ corrected for both (scale ranging from red and/or blue to yellow and/or light blue)). (**a**) Three-dimensional views highlighting in red-yellow the central cluster in the left uncinate fasciculus in which FA was significantly increased in subjects with BD versus controls ($t = 3.0$, $P < 0.05$ corrected). (**b**) Three-dimensional views highlighting in red-yellow an orbitomedial prefrontal cortex cluster in the left uncinate fasciculus in which FA was significantly increased in subjects with BD versus controls ($t = 4.5$, $P < 0.05$ corrected). (**c**) Three-dimensional views highlighting in light blue a cluster in the right uncinate fasciculus in which FA was significantly reduced in subjects with BD versus controls ($t = 3.3$, $P < 0.05$ corrected). MNI, Montreal Neurological Institute (reproduced with permission from Versace *et al.*[64]).

Difficulties are manifold. Mood disorders are highly heterogeneous entities and there is considerable overlap in the statistical distributions between patients with mood disorders and healthy controls in regional brain volumes, receptor BP, BOLD hemodynamic response, blood flow, metabolism and other neuroimaging measures. Thus, unlike other areas of medicine where clinical tests have a clearly defined normal range, in psychiatry there is no consensus on what constitutes an abnormal result on an MRI or PET scan. For instance, there are no standard normative ranges for the volumes of cerebral structures.

Second, neuroimaging techniques—especially fMRI—are highly sensitive to normal temporal fluctuations in patient physiology or to chemical substance intake that may have nothing to do with mood symptoms (for example, caffeine consumption and nicotine),[93,94] medical conditions that are commonly comorbid with mood disorders and may themselves affect imaging data (for example, diabetes mellitus and hypertension),[95] medication, which may independently affect neurophysiology (for example, lithium and antidepressants),[11] and scanner resolution and sensitivity (signal-to-noise), which will limit the type of morphometric and functional changes that can be measured accurately. The outcomes of PET studies are additionally sensitive to the type of radiotracer/ligand administered to the patient and to the statistical methodology used to model BP. The results of some types of scans may also be affected by factors such as smoking status,[96] body mass index[97] and the number of hours of sunshine.[98]

Third, the statistical power of functional imaging methods, and the ability to discriminate WM and GM boundaries using structural MRI, increases with imaging time, potentially leading to a tradeoff between accuracy and time burden/cost. Moreover, PET scanning requires the subject to be exposed to radiation, such that the lowest potentially effective radioactive dose is injected to ensure safety, even though higher injected doses would be needed to optimize the signal-to-noise ratio for some radioligands.

Fourth, in the discussion above, we emphasized the fact that several studies have shown that healthy individuals with a family-history of mood disorders show similar neuroimaging abnormalities to those observed in ill patients. While these endophenotypes can be leveraged to improve our understanding

of the pathophysiology of mood disorders, the existence of these biomarkers poses a serious obstacle for the application of neuroimaging to clinical diagnosis. Specifically, the presence of imaging endophenotypes potentially may decrease diagnostic specificity by increasing the risk for false-positive diagnoses in healthy individuals who share genetic risk factors with depressed relatives.

Fifth, medication is a potent confound not only because it may affect brain structure and function, but also because it may bias classification algorithms. The algorithms may distinguish patients from controls based on the impact of different classes of medication rather than diagnosis-specific neurophysiology. Conversely, if an algorithm is developed on an unmedicated sample, then it may be inaccurate when applied to a medicated subject.

Sixth, the identification of a biomarker that can predict response to a particular type of treatment may be confounded by the placebo effect. In other words, if $\sim 50\%$ of the response to an antidepressant (or other psychiatric) medication is due to the placebo effect,[99,100] that is, a non-specific effect, then this may impede the identification of a biomarker specific for response to the pharmacological agent as opposed to a more general marker of treatment outcome. A significant limitation of most of the extant literature on neuroimaging correlates of treatment response has been the consistent absence of a placebo-control arm (see below).

Classification of mood disorders and/or response to treatment using neuroimaging: empirical evidence

The development of imaging-based diagnostic algorithms that are sufficiently robust to be applied across cohorts and sites will be a significant challenge. Currently, researchers are still in the process of developing robust diagnostic classifiers within just one cohort of patients at a time. The challenge is to determine how best to identify the key prediction signals in the mass of data produced by neuroimaging. One approach is to use sophisticated and powerful statistical techniques such as machine learning. Machine learning refers to a group of statistical methods that are used to develop algorithms to detect patterns or regularities within high-dimensional data. An empirical data training set—for example, the MRI data of DSM-IV-diagnosed patients versus healthy controls—is used to develop an algorithm that optimally distinguishes between these groups. Theoretically, the computer will then be able to make intelligent decisions about new cases based on the examples provided in the training set. That is, the program 'learns' from experience.

Once an algorithm has been developed, the gold standard is to validate the algorithm on an independent cohort. However, as discussed below, the papers published to date have made use of a less stringent validation method—namely the 'leave out one' approach. That is, all subjects except one patient–control pair are initially chosen to comprise the training set and an algorithm that best separates the diagnostic groups from each other is applied to the omitted pair to predict their diagnostic status or treatment response. The process is then iteratively applied to each subject pair to test the ability of the algorithm to distinguish between categories. That is, each omitted subject pair comprises one training example. The 'leave out one' approach is less stringent because one would expect to find significant variation across subject samples. A proportion of this variation is likely to be noise—that is, the confounding effects of temporal fluctuations, medications and other factors discussed above, and a proportion of this variation is likely to result from disease heterogeneity. Only by testing an algorithm on an independent cohort, can one demonstrate that the discriminator is robust to these confounds.

Sun et al.[101] created cortical density maps for 36 healthy controls and 36 patients with recent onset schizophrenia spectrum or affective psychosis. On a group level, the patients displayed reduced GM density in regions such as the anterior cingulate and lateral surfaces of the prefrontal and temporal cortices compared with the control group. Machine learning methods were then applied to the data to test whether these findings could be applied at the individual subject level. Using a sparse multinomial logistic regression classifier, 129 surface voxels were linearly combined for classification allowing for 86% accuracy in distinguishing between patients and controls. Clusters with the highest weightings included the frontal pole, superior and middle temporal regions of the left hemisphere, and the superior temporal, somatomotor and subgenual ACC regions of the right hemisphere.

Fu et al.[102] used the voxel-wise hemodynamic response to sad faces to distinguish acutely depressed patients with MDD ($n = 19$) from healthy controls ($n = 19$) with 82% sensitivity and 89% specificity. Regions with the highest vector weights included the dorsal ACC, middle and superior frontal gyri, hippocampus, caudate, thalamus and amygdala. The same group achieved a less robust 65% sensitivity and 70% specificity with the use of a working memory paradigm in 20 healthy subjects and 20 unmedicated patients with major depression.[103] Interestingly, despite the difference in task paradigm there was some overlap in the regions that distinguished patients and controls in the sad face task—the caudate, and the superior and middle-frontal gyri.

In another study, the hemodynamic response of the default mode and temporal lobe networks during an auditory oddball paradigm was applied a priori to a sample of 14 medicated patients with BD, type I (BD I), 21 medicated patients with schizophrenia, and 26 healthy controls.[104] The authors were able to distinguish BD patients from patients with schizophrenia and healthy controls with 83% sensitivity and 100% specificity. The accuracy of the BD versus healthy control classification was not provided. Most recently, Hahn et al.[105] utilized three independent fMRI paradigms in an attempt to maximize classification accuracy: the passive viewing of emotionally valenced faces, and two different versions of the monetary incentive delay task emphasizing potential winnings and potential losses, respectively. A decision tree algorithm derived from the combination of the imaging task classifiers produced a diagnostic sensitivity of 80% and a specificity of 87% in a sample of 30 patients with depression (both unipolar and bipolar) and 30 healthy controls. The algorithm's ability to distinguish subjects with unipolar depression from BD was not reported.

A Gaussian Process Classifiers machine-based learning approach was recently used to distinguish healthy adolescents with and without a parent with BD from each other with 75% sensitivity and 75% specificity.[106] A discriminating pattern of BOLD activation was found in the superior temporal sulcus and ventromedial PFC when subjects were presented with neutral faces in the context of happy faces. Six out of thirteen of the high-risk adolescents who were followed clinically, subsequently met DSM-IV criteria for MDD or an anxiety disorder. Interestingly, these six individuals had higher Gaussian Process Classifiers risk scores than the seven high-risk subjects who did not become ill.[106] Moreover, three out of the four high-risk subjects that the Gaussian Process Classifiers algorithm incorrectly classified as low-risk, remained healthy at follow-up.[106]

Several studies have recently used machine learning methods to evaluate response to treatment with antidepressant medication. In one such study, a whole-brain voxel-based morphometry analysis predicted treatment response to fluoxetine with 89% sensitivity and 89% specificity. The same algorithm derived from the voxel-based morphometry analysis only differentiated MDD patients ($n = 37$) from healthy controls ($n = 37$) with 65% sensitivity and 70% specificity.[107] Response to treatment was associated with increased GM density of the rostral ACC, left posterior cingulate cortex, left middle frontal gyrus and right occipital cortex at baseline.[107] Gong et al.[108] used structural MRI to

predict antidepressant efficacy in 61 treatment naïve patients with depression. Patients who failed to respond to two adequate trials of an antidepressant were distinguished from treatment responders with 70% sensitivity and 70% specificity based on GM and WM volumes: treatment responders had both greater and lower baseline volumes of different regions in the frontal, temporal, parietal and occipital cortices, as well lower baseline volume of the putamen.[108] Costafreda et al.[109] reported that in 16 unmedicated patients who met criteria for a major depressive episode, pretreatment response to implicitly presented sad faces in regions such as the dorsal ACC, midcingulate gyrus, superior frontal gyrus, and posterior cingulate cortex predicted subsequent response to cognitive behavioral therapy with a sensitivity of 71% and a specificity of 86%.

Other attempts at predicting response to treatment have been less successful. The functional imaging correlates of a verbal working memory task only predicted response to fluoxetine with 52% specificity, although sensitivity was 85%.[103] Conversely, 62% of patients who achieved clinical remission and 75% of patients who did not remit following 8 weeks of antidepressant treatment, were correctly identified as responders and non-responders, respectively, with a sad face processing task.[102]

In sum, current diagnostic and treatment prediction methods have yielded sensitivities and specificities that range from 70 to 90%. That is, ~3 out of 10 patients with a mood disorder would be incorrectly diagnosed as healthy, and ~1 out of 10 healthy individuals would be incorrectly diagnosed with a mood disorder. Nevertheless, none of the above-mentioned studies have achieved this degree of diagnostic success in an independent cohort, and this will be a crucial test for the field. Ultimately, the patient burden and/or risk of the scan, together with its financial cost, will have to be balanced against the potential benefits of testing such as improved outcomes and more cost efficient treatment. The extent to which diagnostic and treatment misclassification will be tolerated by patients, clinicians and the health care industry will ultimately be determined by this cost-benefit ratio.

Independent of the technical challenges involved in developing diagnostic algorithms, we raise the issue of whether the current approach to developing neuroimaging-based tests for the *diagnosis of psychiatric disorders* is philosophically flawed. The claim that the machine learning approach will lead to objective biomarkers of psychiatric illness that will supplant the clinical interview is circular because the algorithms are trained to categorize patients based on clinical (that is, DSM-IV) diagnoses. Yet the *raison d'etre* of the biomarker is the future supersession of the subjective diagnosis as the gold standard. Our current diagnostic categories may subsume multiple distinct disorders and thus attempting to forcibly align neurobiology with DSM diagnoses is arguably regressive.

This view has recently been championed by the National Institute of Mental Health in the form of the Research Domain Criteria initiative which seeks to lay the foundations for a future psychiatric nosology based on neuroscience and genetics rather than clinical observation.[110] The framework for this alternative psychiatric classification system is formed by psychological/ behavioral 'constructs' that are explicitly linked to the underlying neurobiology.[110] For example, in the context of mood disorders, two potentially relevant constructs are 'loss' (HPA axis dysregulation, sustained amygdala reactivity, and so on) and 'response to reward attainment' (reduced activity of the nucleus accumbens, orbitofrontal cortex, and so on). Arguably, imaging-based diagnostic algorithms that differentiate individuals on the basis of Research Domain Criteria constructs such as 'loss' and 'response to reward attainment' potentially would optimize treatment strategies in currently ill patients and allow for the identification of individuals at risk of developing a mood disorder in the future.

REFERENCES

1 De Gruttola VG, Clax P, DeMets DL, Downing GJ, Ellenberg SS, Friedman L et al. Considerations in the evaluation of surrogate endpoints in clinical trials. summary of a National Institutes of Health workshop. Control Clin Trials 2001; 22: 485–502.

2 Mayeux R. Evaluation and use of diagnostic tests in Alzheimer's disease. Neurobiol Aging 1998; 19: 139–143.

3 Consensus report of the Working Group on: "Molecular and Biochemical Markers of Alzheimer's Disease". The Ronald and Nancy Reagan Research Institute of the Alzheimer's Association and the National Institute on Aging Working Group. Neurobiol Aging 1998; 19: 109–116.

4 Clark CM, Schneider JA, Bedell BJ, Beach TG, Bilker WB, Mintun MA et al. Use of florbetapir-PET for imaging beta-amyloid pathology. Jama 2011; 305: 275–283.

5 Lilienfeld AM, Kordan B. A study of variability in the interpretation of chest x-rays in the detection of lung cancer. Cancer Res 1966; 26: 2145–2147.

6 APA. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). APA Press: Washington, D.C., 1994.

7 Savitz JB, Drevets WC. Imaging phenotypes of major depressive disorder: genetic correlates. Neuroscience 2009; 164: 300–330.

8 Hasler G, Northoff G. Discovering imaging endophenotypes for major depression. Mol Psychiatry 2011; 16: 604–619.

9 Delvecchio G, Fossati P, Boyer P, Brambilla P, Falkai P, Gruber O et al. Common and distinct neural correlates of emotional processing in Bipolar Disorder and Major Depressive Disorder: a voxel-based meta-analysis of functional magnetic resonance imaging studies. Eur Neuropsychopharmacol 2012; 22: 100–113.

10 Yuksel C, Ongur D. Magnetic resonance spectroscopy studies of glutamate-related abnormalities in mood disorders. Biol Psychiatry 2010; 68: 785–794.

11 Savitz J, Drevets WC. Bipolar and major depressive disorder: neuroimaging the developmental-degenerative divide. Neurosci Biobehav Rev 2009; 33: 699–771.

12 Savitz JB, Drevets WC. Neuroreceptor imaging in depression. Neurobiol Dis 2013; 52: 49–65.

13 Strakowski SM, Adler CM, Almeida J, Altshuler LL, Blumberg HP, Chang KD et al. The functional neuroanatomy of bipolar disorder: a consensus model. Bipolar Disord 2012; 14: 313–325.

14 Sheline YI, Barch DM, Donnelly JM, Ollinger JM, Snyder AZ, Mintun MA. Increased amygdala response to masked emotional faces in depressed subjects resolves with antidepressant treatment: an fMRI study. Biol Psychiatry 2001; 50: 651–658.

15 Victor TA, Furey ML, Fromm SJ, Ohman A, Drevets WC. Relationship between amygdala responses to masked faces and mood state and treatment in major depressive disorder. Arch Gen Psychiatry 2010; 67: 1128–1138.

16 Victor TA, Furey ML, Fromm SJ, Bellgowan PS, Ohman A, Drevets WC. The extended functional neuroanatomy of emotional processing biases for masked faces in major depressive disorder. PLoS One 2012; 7: e46439.

17 Suslow T, Konrad C, Kugel H, Rumstadt D, Zwitserlood P, Schoning S et al. Automatic mood-congruent amygdala responses to masked facial expressions in major depression. Biol Psychiatry 2010; 67: 155–160.

18 Siegle GJ, Thompson W, Carter CS, Steinhauer SR, Thase ME. Increased amygdala and decreased dorsolateral prefrontal BOLD responses in unipolar depression: related and independent features. Biol Psychiatry 2007; 61: 198–209.

19 Robinson OJ, Cools R, Carlisi CO, Sahakian BJ, Drevets WC. Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. Am J Psychiatry 2012; 169: 152–159.

20 Der-Avakian A, Markou A. The neurobiology of anhedonia and other reward-related deficits. Trends Neurosci 2012; 35: 68–77.

21 Keedwell PA, Andrew C, Williams SC, Brammer MJ, Phillips ML. The neural correlates of anhedonia in major depressive disorder. Biol Psychiatry 2005; 58: 843–853.

22 Epstein J, Pan H, Kocsis JH, Yang Y, Butler T, Chusid J et al. Lack of ventral striatal response to positive stimuli in depressed versus normal subjects. Am J Psychiatry 2006; 163: 1784–1790.

23 Pizzagalli DA, Iosifescu D, Hallett LA, Ratner KG, Fava M. Reduced hedonic capacity in major depressive disorder: evidence from a probabilistic reward task. J Psychiatr Res 2008; 43: 76–87.

24 Treadway MT, Zald DH. Reconsidering anhedonia in depression: lessons from translational neuroscience. *Neurosci Biobehav Rev* 2011; **35**: 537–555.

25 Pizzagalli DA, Holmes AJ, Dillon DG, Goetz EL, Birk JL, Bogdan R et al. Reduced caudate and nucleus accumbens response to rewards in unmedicated individuals with major depressive disorder. *Am J Psychiatry* 2009; **166**: 702–710.

26 Gottesman, II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions *Am J Psychiatry* 2003; **160**: 636–645.

27 Monk CS, Klein RG, Telzer EH, Schroth EA, Mannuzza S, Moulton 3rd JL et al. Amygdala and nucleus accumbens activation to emotional facial expressions in children and adolescents at risk for major depression. *Am J Psychiatry* 2008; **165**: 90–98.

28 Olsavsky AK, Brotman MA, Rutenberg JG, Muhrer EJ, Deveney CM, Fromm SJ et al. Amygdala hyperactivation during face emotion processing in unaffected youth at risk for bipolar disorder. *J Am Acad Child Adolesc Psychiatry* 2012; **51**: 294–303.

29 Gotlib IH, Hamilton JP, Cooney RE, Singh MK, Henry ML, Joormann J. Neural processing of reward and loss in girls at risk for major depression. *Arch Gen Psychiatry* 2010; **67**: 380–387.

30 McCabe C, Woffindale C, Harmer CJ, Cowen PJ. Neural processing of reward and punishment in young people at increased familial risk of depression. *Biol Psychiatry* 2012; **72**: 588–594.

31 Drevets WC, Savitz J, Trimble M. The subgenual anterior cingulate cortex in mood disorders. *CNS Spectr* 2008; **13**: 663–681.

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

33 Hajek T, Kozeny J, Kopecek M, Alda M, Hoschl C. Reduced subgenual cingulate volumes in mood disorders: a meta-analysis. *J Psychiatry Neurosci* 2008; **33**: 91–99.

34 Koolschijn PC, van Haren NE, Lensvelt-Mulders GJ, Hulshoff Pol HE, Kahn RS. Brain volume abnormalities in major depressive disorder: a meta-analysis of magnetic resonance imaging studies. *Hum Brain Mapp* 2009; **30**: 3719–3735.

35 Arnone D, McIntosh AM, Ebmeier KP, Munafo MR, Anderson IM. Magnetic resonance imaging studies in unipolar depression: systematic review and meta-regression analyses. *Eur Neuropsychopharmacol* 2012; **22**: 1–16.

36 Kempton MJ, Salvador Z, Munafo MR, Geddes JR, Simmons A, Frangou S et al. Structural neuroimaging studies in major depressive disorder. Meta-analysis and comparison with bipolar disorder. *Arch Gen Psychiatry* 2011; **68**: 675–690.

37 Hafeman DM, Chang KD, Garrett AS, Sanders EM, Phillips ML. Effects of medication on neuroimaging findings in bipolar disorder: an updated review. *Bipolar Disord* 2012; **14**: 375–410.

38 Moore GJ, Bebchuk JM, Wilds IB, Chen G, Manji HK. Lithium-induced increase in human brain grey matter. *Lancet* 2000; **356**: 1241–1242.

39 Moore GJ, Cortese BM, Glitz DA, Zajac-Benitez C, Quiroz JA, Uhde TW et al. A longitudinal study of the effects of lithium treatment on prefrontal and subgenual prefrontal gray matter volume in treatment-responsive bipolar disorder patients. *J Clin Psychiatry* 2009; **70**: 699–705.

40 Savitz J, Nugent AC, Bogers W, Liu A, Sills R, Luckenbaugh DA et al. Amygdala volume in depressed patients with bipolar disorder assessed using high resolution 3T MRI: the impact of medication. *Neuroimage* 2010; **49**: 2966–2976.

41 Savitz JB, Nugent AC, Bogers W, Roiser JP, Bain EE, Neumeister A et al. Habenula volume in bipolar disorder and major depressive disorder: a high-resolution magnetic resonance imaging study. *Biol Psychiatry* 2011; **69**: 336–343.

42 Chen MC, Hamilton JP, Gotlib IH. Decreased hippocampal volume in healthy girls at risk of depression. *Arch Gen Psychiatry* 2010; **67**: 270–276.

43 Carballedo A, Lisiecka D, Fagan A, Saleh K, Ferguson Y, Connolly G et al. Early life adversity is associated with brain changes in subjects at family risk for depression. *World J Biol Psychiatry* 2012; **13**: 569–578.

44 Boes AD, McCormick LM, Coryell WH, Nopoulos P. Rostral anterior cingulate cortex volume correlates with depressed mood in normal healthy children. *Biol Psychiatry* 2008; **63**: 391–397.

45 Glahn DC, Curran JE, Winkler AM, Carless MA, Kent Jr. JW, Charlesworth JC et al. High dimensional endophenotype ranking in the search for major depression risk genes. *Biol Psychiatry* 2012; **71**: 6–14.

46 McDonald C, Bullmore ET, Sham PC, Chitnis X, Wickham H, Bramon E et al. Association of genetic risks for schizophrenia and bipolar disorder with specific and generic brain structural endophenotypes. *Arch Gen Psychiatry* 2004; **61**: 974–984.

47 Stockmeier CA, Mahajan GJ, Konick LC, Overholser JC, Jurjus GJ, Meltzer HY et al. Cellular changes in the postmortem hippocampus in major depression. *Biol Psychiatry* 2004; **56**: 640–650.

48 Ongur D, Drevets WC, Price JL. Glial reduction in the subgenual prefrontal cortex in mood disorders. *Proc Natl Acad Sci USA* 1998; **95**: 13290–13295.

49 Price JL, Drevets WC. Neurocircuitry of mood disorders. *Neuropsychopharmacology* 2010; **35**: 192–216.

50 Taylor MJ, Selvaraj S, Norbury R, Jezzard P, Cowen PJ. Normal glutamate but elevated myo-inositol in anterior cingulate cortex in recovered depressed patients. *J Affect Disord* 2009; **119**: 186–189.

51 Hasler G, van der Veen JW, Tumonis T, Meyers N, Shen J, Drevets WC. Reduced prefrontal glutamate/glutamine and gamma-aminobutyric acid levels in major depression determined using proton magnetic resonance spectroscopy. *Arch Gen Psychiatry* 2007; **64**: 193–200.

52 Sanacora G, Mason GF, Rothman DL, Behar KL, Hyder F, Petroff OA et al. Reduced cortical gamma-aminobutyric acid levels in depressed patients determined by proton magnetic resonance spectroscopy. *Arch Gen Psychiatry* 1999; **56**: 1043–1047.

53 Gabbay V, Mao X, Klein RG, Ely BA, Babb JS, Panzer AM et al. Anterior cingulate cortex gamma-aminobutyric acid in depressed adolescents: relationship to anhedonia. *Arch Gen Psychiatry* 2012; **69**: 139–149.

54 Sheline YI, Barch DM, Price JL, Rundle MM, Vaishnavi SN, Snyder AZ et al. The default mode network and self-referential processes in depression. *Proc Natl Acad Sci USA* 2009; **106**: 1942–1947.

55 Northoff G, Wiebking C, Feinberg T, Panksepp J. The 'resting-state hypothesis' of major depressive disorder-a translational subcortical-cortical framework for a system disorder. *Neurosci Biobehav Rev* 2011; **35**: 1929–1945.

56 Sheline YI, Price JL, Yan Z, Mintun MA. Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. *Proc Natl Acad Sci USA* 2010; **107**: 11020–11025.

57 Greicius MD, Flores BH, Menon V, Glover GH, Solvason HB, Kenna H et al. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol Psychiatry* 2007; **62**: 429–437.

58 Pizzagalli DA. Frontocingulate dysfunction in depression: toward biomarkers of treatment response. *Neuropsychopharmacology* 2011; **36**: 183–206.

59 Alexopoulos GS, Meyers BS, Young RC, Campbell S, Silbersweig D, Charlson M. 'Vascular depression' hypothesis. *Arch Gen Psychiatry* 1997; **54**: 915–922.

60 Krishnan KR, Hays JC, Blazer DG. MRI-defined vascular depression. *Am J Psychiatry* 1997; **154**: 497–501.

61 Botteron KN, Vannier MW, Geller B, Todd RD, Lee BC. Preliminary study of magnetic resonance imaging characteristics in 8- to 16-year-olds with mania. *J Am Acad Child Adolesc Psychiatry* 1995; **34**: 742–749.

62 McDonald WM, Tupler LA, Marsteller FA, Figiel GS, DiSouza S, Nemeroff CB et al. Hyperintense lesions on magnetic resonance images in bipolar disorder. *Biol Psychiatry* 1999; **45**: 965–971.

63 Lyoo IK, Lee HK, Jung JH, Noam GG, Renshaw PF. White matter hyperintensities on magnetic resonance imaging of the brain in children with psychiatric disorders. *Compr Psychiatry* 2002; **43**: 361–368.

64 Versace A, Almeida JR, Hassel S, Walsh ND, Novelli M, Klein CR et al. Elevated left and reduced right orbitomedial prefrontal fractional anisotropy in adults with bipolar disorder revealed by tract-based spatial statistics. *Arch Gen Psychiatry* 2008; **65**: 1041–1052.

65 Beyer JL, Taylor WD, MacFall JR, Kuchibhatla M, Payne ME, Provenzale JM et al. Cortical white matter microstructural abnormalities in bipolar disorder. *Neuropsychopharmacology* 2005; **30**: 2225–2229.

66 Adler CM, Holland SK, Schmithorst V, Wilke M, Weiss KL, Pan H et al. Abnormal frontal white matter tracts in bipolar disorder: a diffusion tensor imaging study. *Bipolar Disord* 2004; **6**: 197–203.

67 Houenou J, Wessa M, Douaud G, Leboyer M, Chanraud S, Perrin M et al. Increased white matter connectivity in euthymic bipolar patients: diffusion tensor tractography between the subgenual cingulate and the amygdalo-hippocampal complex. *Mol Psychiatry* 2007; **12**: 1001–1010.

68 Lu LH, Zhou XJ, Fitzgerald J, Keedy SK, Reilly JL, Passarotti AM et al. Microstructural abnormalities of white matter differentiate pediatric and adult-onset bipolar disorder. *Bipolar Disord* 2012; **14**: 597–606.

69 Vederine FE, Wessa M, Leboyer M, Houenou J. A meta-analysis of whole-brain diffusion tensor imaging studies in bipolar disorder. *Prog Neuropsychopharmacol Biol Psychiatry* 2011; **35**: 1820–1826.

70 Liao Y, Huang X, Wu Q, Yang C, Kuang W, Du M et al. Is depression a disconnection syndrome? Meta-analysis of diffusion tensor imaging studies in patients with MDD. *J Psychiatry Neurosci* 2012; **37**: 110180.

71 Zhu X, Wang X, Xiao J, Zhong M, Liao J, Yao S. Altered white matter integrity in first-episode, treatment-naive young adults with major depressive disorder: a tract-based spatial statistics study. *Brain Res* 2011; **1369**: 223–229.

72 Cullen KR, Klimes-Dougan B, Muetzel R, Mueller BA, Camchong J, Houri A et al. Altered white matter microstructure in adolescents with major depression: a preliminary study. *J Am Acad Child Adolesc Psychiatry* 2010; **49**: 173–183 e171.

73 Anand A, Li Y, Wang Y, Wu J, Gao S, Bukhari L et al. Activity and connectivity of brain mood regulating circuit in depression: a functional magnetic resonance study. *Biol Psychiatry* 2005; **57**: 1079–1088.

74 Chen CH, Suckling J, Ooi C, Fu CH, Williams SC, Walsh ND et al. Functional coupling of the amygdala in depressed patients treated with antidepressant medication. Neuropsychopharmacology 2008; 33: 1909–1918.

75 Gulseren S, Gurcan M, Gulseren L, Gelal F, Erol A. T2 hyperintensities in bipolar patients and their healthy siblings. Arch Med Res 2006; 37: 79–85.

76 Gunde E, Novak T, Kopecek M, Schmidt M, Propper L, Stopkova P et al. White matter hyperintensities in affected and unaffected late teenage and early adulthood offspring of bipolar parents: a two-center high-risk study. J Psychiatr Res 2011; 45: 76–82.

77 Frazier JA, Breeze JL, Papadimitriou G, Kennedy DN, Hodge SM, Moore CM et al. White matter abnormalities in children with and at risk for bipolar disorder. Bipolar Disord 2007; 9: 799–809.

78 Versace A, Ladouceur CD, Romero S, Birmaher B, Axelson DA, Kupfer DJ et al. Altered development of white matter in youth at high familial risk for bipolar disorder: a diffusion tensor imaging study.. J Am Acad Child Adolesc Psychiatry 2010; 49: 1249–1259 1259 e1241.

79 Mahon K, Burdick KE, Ikuta T, Braga RJ, Gruner P, Malhotra AK et al. Abnormal temporal lobe white matter as a biomarker for genetic risk of bipolar disorder. Biol Psychiatry 2013; 73: 177–182.

80 Sprooten E, Sussmann JE, Clugston A, Peel A, McKirdy J, Moorhead TW et al. White matter integrity in individuals at high genetic risk of bipolar disorder. Biol Psychiatry 2011; 70: 350–356.

81 Keedwell PA, Chapman R, Christiansen K, Richardson H, Evans J, Jones DK. Cingulum white matter in young women at risk of depression: the effect of family history and anhedonia. Biol Psychiatry 2012; 72: 296–302.

82 Huang H, Fan X, Williamson DE, Rao U. White matter changes in healthy adolescents at familial risk for unipolar depression: a diffusion tensor imaging study. Neuropsychopharmacology 2011; 36: 684–691.

83 Hirvonen J, Karlsson H, Kajander J, Lepola A, Markkula J, Rasi-Hakala H et al. Decreased brain serotonin 5-HT1A receptor availability in medication-naive patients with major depressive disorder: an in-vivo imaging study using PET and [carbonyl-11C]WAY-100635. Int J Neuropsychopharmacol 2008; 11: 465–476.

84 Bhagwagar Z, Rabiner EA, Sargent PA, Grasby PM, Cowen PJ. Persistent reduction in brain serotonin1A receptor binding in recovered depressed men measured by positron emission tomography with [11C]WAY-100635. Mol Psychiatry 2004; 9: 386–392.

85 Drevets WC, Frank E, Price JC, Kupfer DJ, Greer PJ, Mathis C. Serotonin type-1A receptor imaging in depression. Nucl Med Biol 2000; 27: 499–507.

86 Drevets WC, Thase ME, Moses-Kolko EL, Price J, Frank E, Kupfer DJ et al. Serotonin-1A receptor imaging in recurrent depression: replication and literature review. Nucl Med Biol 2007; 34: 865–877.

87 Sargent PA, Kjaer KH, Bench CJ, Rabiner EA, Messa C, Meyer J et al. Brain serotonin1A receptor binding measured by positron emission tomography with [11C]WAY-100635: effects of depression and antidepressant treatment. Arch Gen Psychiatry 2000; 57: 174–180.

88 Drevets WC, Frank E, Price JC, Kupfer DJ, Holt D, Greer PJ et al. PET imaging of serotonin 1A receptor binding in depression. Biol Psychiatry 1999; 46: 1375–1387.

89 Cannon DM, Ichise M, Rollis D, Klaver JM, Gandhi SK, Charney DS et al. Elevated serotonin transporter binding in major depressive disorder assessed using positron emission tomography and [11C]DASB; comparison with bipolar disorder. Biol Psychiatry 2007; 62: 870–877.

90 Cannon DM, Ichise M, Fromm SJ, Nugent AC, Rollis D, Gandhi SK et al. Serotonin transporter binding in bipolar disorder assessed using [11C]DASB and positron emission tomography. Biol Psychiatry 2006; 60: 207–217.

91 Parsey RV, Olvet DM, Oquendo MA, Huang YY, Ogden RT, Mann JJ. Higher 5-HT1A receptor binding potential during a major depressive episode predicts poor treatment response: preliminary data from a naturalistic study. Neuropsychopharmacology 2006; 31: 1745–1749.

92 Parsey RV, Hastings RS, Oquendo MA, Huang YY, Simpson N, Arcement J et al. Lower serotonin transporter binding potential in the human brain during major depressive episodes. Am J Psychiatry 2006; 163: 52–58.

93 Chen Y, Parrish TB. Caffeine's effects on cerebrovascular reactivity and coupling between cerebral blood flow and oxygen metabolism. Neuroimage 2009; 44: 647–652.

94 Hahn B, Ross TJ, Yang Y, Kim I, Huestis MA, Stein EA. Nicotine enhances visuospatial attention by deactivating areas of the resting brain default network. J Neurosci 2007; 27: 3477–3489.

95 Bruehl H, Wolf OT, Sweat V, Tirsi A, Richardson S, Convit A. Modifiers of cognitive function and brain structure in middle-aged and elderly individuals with type 2 diabetes mellitus. Brain Res 2009; 1280: 186–194.

96 Meyer JH, McNeely HE, Sagrati S, Boovariwala A, Martin K, Verhoeff NP et al. Elevated putamen D(2) receptor binding potential in major depression with motor retardation: an [11C]raclopride positron emission tomography study. Am J Psychiatry 2006; 163: 1594–1602.

97 Erritzoe D, Frokjaer VG, Haahr MT, Kalbitzer J, Svarer C, Holst KK et al. Cerebral serotonin transporter binding is inversely related to body mass index. Neuroimage 2010; 52: 284–289.

98 Praschak-Rieder N, Willeit M, Wilson AA, Houle S, Meyer JH. Seasonal variation in human brain serotonin transporter binding. Arch Gen Psychiatry 2008; 65: 1072–1078.

99 Walsh BT, Seidman SN, Sysko R, Gould M. Placebo response in studies of major depression: variable, substantial, and growing. Jama 2002; 287: 1840–1847.

100 Rief W, Nestoriuc Y, Weiss S, Welzel E, Barsky AJ, Hofmann SG. Meta-analysis of the placebo response in antidepressant trials. J Affect Disord 2009; 118: 1–8.

101 Sun D, van Erp TG, Thompson PM, Bearden CE, Daley M, Kushan L et al. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. Biol Psychiatry 2009; 66: 1055–1060.

102 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. Biol Psychiatry 2008; 63: 656–662.

103 Marquand AF, Mourao-Miranda J, Brammer MJ, Cleare AJ, Fu CH. Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. Neuroreport 2008; 19: 1507–1511.

104 Calhoun VD, Maciejewski PK, Pearlson GD, Kiehl KA. Temporal lobe and "default" hemodynamic brain modes discriminate between schizophrenia and bipolar disorder. Hum Brain Mapp 2008; 29: 1265–1275.

105 Hahn T, Marquand AF, Ehlis AC, Dresler T, Kittel-Schneider S, Jarczok TA et al. Integrating neurobiological markers of depression. Arch Gen Psychiatry 2011; 68: 361–368.

106 Mourao-Miranda J, Oliveira L, Ladouceur CD, Marquand A, Brammer M, Birmaher B et al. Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. PLoS One 2012; 7: e29482.

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

108 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: 1497–1503.

109 Costafreda SG, Khanna A, Mourao-Miranda J, Fu CH. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. Neuroreport 2009; 20: 637–641.

110 Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry 2010; 167: 748–751.

111 Savitz JB, Bonne O, Nugent AC, Vythilingam M, Bogers W, Charney DS et al. Habenula volume in post-traumatic stress disorder measured with high-resolution MRI. Biol Mood Anxiety Disord 2011; 1: 7.