Archival Report

Biological Psychiatry

Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies

Joseph Kambeitz, Carlos Cabral, Matthew D. Sacchet, Ian H. Gotlib, Roland Zahn, Mauricio H. Serpa, Martin Walter, Peter Falkai, and Nikolaos Koutsouleris

ABSTRACT

BACKGROUND: Multiple studies have examined functional and structural brain alteration in patients diagnosed with major depressive disorder (MDD). The introduction of multivariate statistical methods allows investigators to utilize data concerning these brain alterations to generate diagnostic models that accurately differentiate patients with MDD from healthy control subjects (HCs). However, there is substantial heterogeneity in the reported results, the methodological approaches, and the clinical characteristics of participants in these studies.

METHODS: We conducted a meta-analysis of all studies using neuroimaging (volumetric measures derived from T1-weighted images, task-based functional magnetic resonance imaging [MRI], resting-state MRI, or diffusion tensor imaging) in combination with multivariate statistical methods to differentiate patients diagnosed with MDD from HCs.

RESULTS: Thirty-three ($k = 33$) samples including 912 patients with MDD and 894 HCs were included in the meta-analysis. Across all studies, patients with MDD were separated from HCs with 77% sensitivity and 78% specificity. Classification based on resting-state MRI (85% sensitivity, 83% specificity) and on diffusion tensor imaging data (88% sensitivity, 92% specificity) outperformed classifications based on structural MRI (70% sensitivity, 71% specificity) and task-based functional MRI (74% sensitivity, 77% specificity).

CONCLUSIONS: Our results demonstrate the high representational capacity of multivariate statistical methods to identify neuroimaging-based biomarkers of depression. Future studies are needed to elucidate whether multivariate neuroimaging analysis has the potential to generate clinically useful tools for the differential diagnosis of affective disorders and the prediction of both treatment response and functional outcome.

Keywords: Affective disorder, Classification, Diagnosis, Prediction, Sensitivity, Specificity

http://dx.doi.org/10.1016/j.biopsych.2016.10.028

Major depressive disorder (MDD) has a lifetime prevalence of 14.6%, making it one of the most common psychiatric disorders worldwide (1). Reliable diagnosis of MDD is a primary prerequisite for effective pharmacological and psychological interventions (2). Currently, the diagnosis of depression is based on the phenomenological evaluation of symptoms and behavior by trained clinicians. Scientists have posited that neuroimaging holds "diagnostic potential" given findings in multiple studies of significant anomalies in brain structure (3–5), function (6–8), and neurochemistry (9,10) in patients with depression. Even though these meta-analyses indicate that brain changes are replicable across studies, the alterations are often small and do not allow a reliable differentiation between patients and control subjects (11). Thus, neuroimaging markers are not included in clinical practice to guide decisions concerning psychiatric diagnosis (12,13). This might result from the higher costs associated with neuroimaging examinations. Moreover, most of the previous neuroimaging studies in MDD have taken a univariate approach, which has important consequences in terms of the clinical applicability of the obtained results. For example, univariate approaches neglect the highly interconnected nature of the brain and, consequently, the statistical dependency of the given units of analysis (e.g., voxels or regions of interest) (14). Moreover, even if two groups (e.g., patients with depression and healthy control subjects [HCs]) differ at a statistically significant level with respect to a target variable (e.g., hippocampal volume), there is typically substantial overlap of the two distributions, hindering reliable differentiation of depressed from nondepressed individuals.

To address these limitations, investigators have begun to apply multivariate statistical methods to the analysis of neuroimaging data (15,16). By focusing on patterns of brain changes that are distributed across multiple regions, these methods allow for the generation of statistical models with high diagnostic or predictive power. In this context, a recent meta-analysis showed that patients with schizophrenia can be accurately differentiated from healthy volunteers in 80% of the cases using only neuroimaging-based diagnostic models (17). Moreover, these methods may facilitate the development

SEE COMMENTARY ON PAGE 306

 ISSN: 0006-3223

of neuroimaging tools to distinguish among different psychiatric disorders (18–21) or to predict clinical outcomes (22–24). Indeed, multiple proof-of-concept studies have successfully used multivariate statistical methods to guide the diagnosis of depression based on structural magnetic resonance imaging (sMRI) data (19,21,25–27), resting-state functional MRI (rsfMRI) data (26,28–34), and task-based functional MRI (fMRI) data (35–39). The sensitivity and the specificity reported in these studies both range from 70% to 90%. This variable diagnostic performance may be due to methodological differences among these studies with respect to the neuroimaging data modality, preprocessing protocol, classification algorithm, or the cross-validation (CV) procedure used. In addition, these studies differ with respect to demographic and clinical characteristics of depressed patients. Differences in performance and study heterogeneity make it difficult to evaluate the potential of neuroimaging to identify diagnostic biomarkers for depression. Here, we report the results of a meta-analysis conducted on studies that used multivariate statistical methods to differentiate patients with depression from HCs. This meta-analytic approach allows us to quantify the ability of multivariate methods to identify depression-related patterns in neuroimaging data. In this way, we investigate the neurobiological construct validity of the current clinical definition of MDD.

METHODS AND MATERIALS

Search and Study Selection Strategy

We searched the electronic PubMed database from January 1, 1950, up to June 31, 2015 (see the Supplement for details). Subsequently, we screened studies according to the following criteria: To be included in the meta-analysis, a paper needed to report results of a neuroimaging-based, supervised, multivariate two-group classification model separating MDD patients from HCs. Studies were included if the following measures of classification performance were available or if data allowed for the calculation of the following parameters: true positives (TP), true negatives (TN), false positives (FP), false negatives (FN). In cases in which insufficient data were reported, the authors were asked to provide additional information regarding their published reports. The results of the literature search are presented in a flowchart following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (40) (see the Supplement and Supplemental Figure S1).

Data Extraction

The main outcome was the diagnostic accuracy of the multivariate diagnostic models when applied to patients with MDD and HCs as indicated by sensitivity [= TP / (TP + FN)] and specificity [= TN / (TN + FP)]. Additional information was extracted from the selected studies as follows: names of the authors, year of publication, demographic characteristics of HC and patient groups [group size, age, sex, medication status, symptoms as measured by the Hamilton Depression Rating Scale (HAMD) (41) or the Beck Depression Inventory (42)], neuroimaging modality (volumetric measures derived of T1-weighted MRI images sMRI, task-based fMRI, rsfMRI, positron emission tomography, single photon emission computed tomography, diffusion tensor imaging [DTI], scanner type, image resolution), characteristics of the neuroimaging preprocessing, configuration of the classification algorithm, and type of the cross-validation procedure (e.g., leave-one-out, *k*-fold cross-validation). To ensure accuracy of data extraction, two authors separately performed extraction and disagreements were resolved in a consensus conference.

Data Analysis

In the present analysis we implemented a random-effects, bivariate meta-analytical model as introduced by Reitsma *et al.* (43). Results of the meta-analysis are presented in forest plots separately for sensitivity and specificity. Summary estimates for sensitivity and specificity are provided separately for sMRI, task-based fMRI, rsfMRI, or DTI studies, and for all studies combined. The robustness of the results and the effects of potentially confounding variables (e.g., age, sex ratio, year of publication) were investigated by adding moderator variables to the bivariate regression model. Furthermore, we tested for differences between studies in the clinical variables using univariate analysis of variance. Publication bias was assessed by creating funnel plots by plotting log diagnostic odds ratios (logDORs) for all studies against $\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ with n_1 and n_2 representing the sample sizes of the patient and the HC group, respectively. This measure is proportional to the inverted square root of the effective sample size (ESS): $\frac{1}{\sqrt{ESS}}$. In case of a publication bias, the distribution of studies in the funnel plot is asymmetrical. A statistical test for funnel plot asymmetry is provided by a regression of logDOR with $\frac{1}{\sqrt{ESS}}$ weighted by ESS (44). As an exploratory analysis, we generated a multivariate regression model using the elastic net algorithm to predict logDOR of individual studies based on 23 clinical and methodological variables (see the Supplement for details). All computations were performed using the *R* statistical programming language version 3.3.1 (45) with the packages mada (46) and glmnet (47).

RESULTS

Meta-analysis

The initial literature search identified 641 studies of interest. After screening all studies and applying the inclusion criteria, 608 studies were excluded (see the Supplement and Supplemental Figure S1 for a flowchart of the literature search). The final sample consisted of 33 studies with a total of 912 patients (mean age: 34.27 years) and 894 HCs (mean age: 32.81 years). From those studies, 14 samples used sMRI (19–21,25–27,48–54), 9 samples used rsfMRI (26,29,31,33,54–58), 9 samples used fMRI (35–37,39,59–63), and 6 samples used DTI (26,64–67) to build predictive models (see the Supplement and Supplemental Table S1 for an overview of the characteristics of the studies; please note that some studies provide more than one sample). There were no studies available using single photon emission computed tomography methodology. One study reported 85% classification accuracy using $^{[18]}$fluorodeoxyglucose positron emission tomography but was excluded from further analysis due to the small number of available studies (68).

Meta-analysis of all studies indicated a sensitivity of 76.66% (95% confidence interval [CI]: 71.95%–80.80%) and a specificity of 77.7% (95% CI: 73.7%–81.35%) (see Figures 1 and 2). Visual inspection of funnel plots and regression test for funnel plot asymmetry ($p = .69$) did not indicate the presence of a publication bias (see the Supplement and Supplemental Figure S2). Moreover, there was no relationship between size of the investigated samples and sensitivity of specificity ($p > .1$, see the Supplement and Supplemental Figure S4). Different neuroimaging modalities (sMRI, fMRI, DTI, rsfMRI) were compared using a moderator analysis. Resting-state MRI studies, compared with sMRI studies, showed higher sensitivity ($p = .007$) and specificity ($p = .017$). There were no significant differences compared with DTI or fMRI studies (all $p > .05$). DTI studies showed a higher sensitivity ($p = .017$) and specificity ($p = .006$) than did sMRI studies, but not fMRI studies ($p > .05$, see Figure 3A).

Subanalysis for every neuroimaging modality showed the following results (see Table 1, Figures 1 and 2). For the subsample of sMRI studies, there was a sensitivity of 69.85% (95% CI: 61.81%–76.83%) (Figure 1) and a specificity of 71.13% (95% CI: 65.41%–76.25%) (see Figure 2). For the subsample of task-related fMRI studies, there was a sensitivity of 74.06% (95% CI: 67.17%–79.94%) (see Figure 1) and a specificity of 77.20% (95% CI: 69.92%–83.15%) (see Figure 2). For the subsample of rsfMRI studies, there was a sensitivity of 85.39% (95% CI: 74.75%–92.02%) (see Figure 1) and a specificity of 82.59% (95% CI: 74.64%–88.43%) (see Figure 2). For the subsample of DTI studies, there was sensitivity of 88.16% (95% CI: 74.18%–95.07%) (see Figure 1) and a specificity of 91.51% (95% CI: 97.15%–77.32%) (see Figure 2). Visual inspection of funnel plots and regression tests for funnel plot asymmetry did not indicate presence of publication bias in the meta-analysis of studies using sMRI ($p = .97$), DTI ($p = .68$), fMRI ($p = .64$), or rsfMRI ($p = .25$).

There was no significant effect of HAMD score on sensitivity or specificity ($p > .80$) (Figure 3C). There was no effect of participants' age on sensitivity ($p = .112$) or specificity ($p = .476$) (see Figure 3B) in the whole sample including all neuroimaging

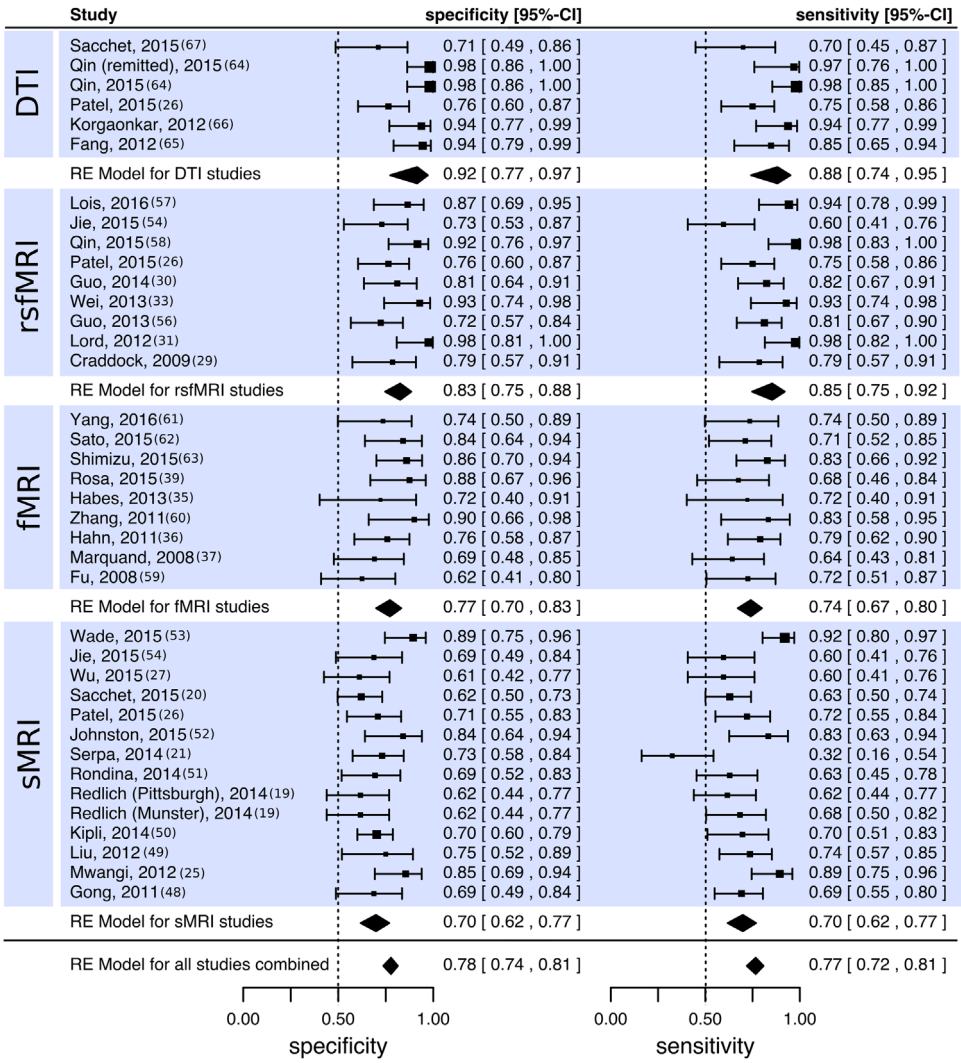

Figure 1. Forest plot of sensitivities and specificities. Summary estimates for sensitivity are computed using the approach described by Reitsma *et al.* (43). CI, confidence interval; DTI, diffusion tensor imaging; fMRI, functional magnetic resonance imaging; RE, random effects; rsfMRI, resting-state functional magnetic resonance imaging; sMRI, structural magnetic resonance imaging.

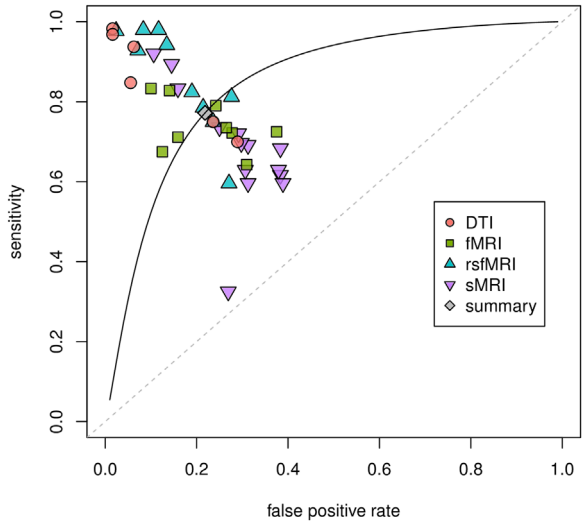

Figure 2. Summary receiver operating characteristic curve of the Reitsma model with the summary sensitivity and false positive rate indicated in black as well as color-coded sensitivity and false positive rates of the individual studies of different imaging modalities. DTI, diffusion tensor imaging; fMRI, functional magnetic resonance imaging; rsfMRI, resting-state functional magnetic resonance imaging; sMRI, structural magnetic resonance imaging.

modalities. Similarly, there was no effect of participants' age in the fMRI or sMRI studies. There was a significant effect of age on sensitivity in the DTI studies ($p = .014$) and in the rsfMRI studies ($p = .015$) but not on specificity ($p > .05$). Sex ratio of patients and of HCs was not related to sensitivity ($p = .414$ and $p = .302$, respectively) or specificity ($p = .582$ and $p = .776$, respectively).

There were heterogeneous methodological approaches in the studies included in the present meta-analysis. When comparing diagnostic accuracies between different cross-validation schemes, there was a higher sensitivity in studies employing twofold CV as compared with leave-one-out and leave-one-subject-per-group-out CV ($p = .035$ and $p = .020$, respectively), but no differences in specificity (see the Supplement and Supplemental Figure S5). Because of the heterogeneous methodological approaches of the studies included in the present meta-analysis, we could not make a statistically valid comparison to test the effects of different classification algorithms on classification accuracy. Thus, we provide a descriptive overview of classification performance associated with different algorithms (see the Supplement and Supplemental Figure S2).

As an exploratory analysis, we tested whether model performance (logDOR) could be predicted on the basis of clinical and methodological variables of the individual studies. Our results indicate that predicted logDOR correlated with true logDOR with $r = .44$ ($p = .002$). The five most important variables in this meta-learning model (as measured by the absolute value of their coefficient averaged across outer-CV folds) were "depression severity: severe," "feature selection: filter," and "patients (age)" (all associated with higher logDOR), as well as "data: sMRI" and "feature selection: none" (all associated with lower logDOR; see the Supplement for further details).

DISCUSSION

We present meta-analyses of a total of 33 studies with a total of 912 patients diagnosed with MDD and 894 HCs. Across all studies, neuroimaging-based diagnostic models were able to differentiate patients from HCs with 77% sensitivity and 78% specificity. These results were robust with respect to potential confounding variables such as age of patients and control subjects, sex ratio, and year of publication. There was no evidence for a publication bias. Resting-state fMRI studies (85% sensitivity, 83% specificity) and DTI studies (88% sensitivity, 92% specificity) outperformed sMRI (70% sensitivity, 71% specificity) and task-based fMRI studies (74% sensitivity, 77% specificity).

Different Neuroimaging Modalities

Our results suggest superior classification accuracy of diagnostic models based on rsfMRI or DTI data, compared with sMRI or task-based fMRI data. It is noteworthy that in a previous analysis, we found rsfMRI to outperform other neuroimaging modalities in differentiating patients with schizophrenia from HCs (17). We should note, however, that a limited number of studies were available using these neuroimaging modalities, so the results need to be interpreted with caution. If this pattern can be confirmed in future analyses, it suggests that rsfMRI and DTI data are the most informative neuroimaging metrics when classifying patients with psychiatric diagnoses versus HCs. On one hand, there might be factors driving this effect that are related to technical details of the neuroimaging methodology. For example, DTI and rsfMRI use scan sequences that typically take longer amounts of time to acquire than do structural MRI sequences and, therefore, might be more susceptible to motion artifacts (69–71).

Table 1. Results from Bivariate Meta-analyses Applying the Approach by Reitsma *et al.* (43)

Data	Control Subjects (*n*)	Patients (*n*)	Sensitivity	Specificity	Positive LR	Negative LR	Diagnostic OR
sMRI	482	450	69.85 (61.81–76.83)	71.13 (65.41–76.25)	2.44 (1.85–3.14)	0.428 (0.312–0.566)	5.93 (3.31–9.87)
fMRI	179	183	74.06 (67.17–79.94)	77.2 (69.92–83.15)	3.29 (2.41–4.45)	0.339 (0.256–0.436)	10.00 (5.77–16.2)
rsfMRI	237	243	85.39 (74.75–92.02)	82.59 (74.64–88.43)	5.03 (3.07–7.76)	0.186 (0.0925–0.326)	31.90 (9.79–78.2)
DTI	162	135	88.16 (74.18–95.07)	91.51 (77.32–97.15)	12.30 (3.42–32.8)	0.145 (0.0513–0.32)	133.00 (11.2–573)
All combined	1060	1011	76.66 (71.95–80.8)	77.76 (73.7–81.35)	3.47 (2.79–4.26)	0.302 (0.24–0.374)	11.70 (7.57–17.4)

Values are summary estimates of random-effects models (95% confidence interval). Positive LR, negative LR, and diagnostic ORs are estimated via Markov chain Monte Carlo (89).

DTI, diffusion tensor imaging; fMRI, functional magnetic resonance imaging; LR, likelihood ratio; OR, odds ratio; rsfMRI, resting-state functional magnetic resonance imaging; sMRI, structural magnetic resonance imaging.

If motion is related to psychiatric diagnosis, then these artifacts might in turn be informative for psychiatric classification and could be picked up by the multivariate classification algorithm. The lower performance of classifiers based on task-based fMRI data compared with DTI and rsfMRI data might be caused by the lower test-retest reliability of this method (72), the dependency on cognitive performance on this task, habituation while performing the task, or variable degree of validity of the employed paradigms for the pathology of MDD. Depending on the preprocessing and the feature selection procedures employed in the task-based fMRI studies, these effects might add noise to the recorded data, which, in turn, might reduce the discriminative power of extracted measures for the subsequent classification. Alternatively, rsfMRI and DTI may capture brain alterations that are more predictive in the context of classifying MDD. It is noteworthy that whereas both modalities are often used to investigate brain connectivity, rsfMRI is a functional measure and DTI is structural. This inherent difference suggests that these modalities capture complementary aspects of the neuropathology of MDD and, thus, that they could be combined in a multimodal classification model to improve performance. To date, there is only one study that compared unimodal with multimodal classification approaches for MDD (26). In that study, multimodal classification (70% accuracy) was outperformed by unimodal classification based on DTI (77%), rsfMRI (77%), and T2 images (77%). Of note, Patel *et al.* (26) assessed a sample of subjects with late-life depression, and thus these results may not generalize to individual subjects with first depressive episodes that usually occur at age 30 (73). Moreover, there is evidence from studies of other psychiatric disorders such as schizophrenia indicating that multimodal classification, when compared with unimodal approaches, improves accuracy (74,75).

Different Classification Algorithms and CV Schemes

The vast majority of studies in the present analysis used a support-vector machine (SVM) algorithm to classify patients (~60%). Looking at neuroimaging modalities separately, SVM was the most frequently used algorithm for DTI (~83%), fMRI (57%), rsfMRI (75%), and sMRI studies (~76%). Some studies used a Gaussian-process classifier (19,36), neural networks (30), random forests (50), *k*-means (50), random trees (50), or decision trees (26). Only two studies systematically investigated different classification algorithms within the same sample (19,26). Redlich *et al*. (19) reported higher accuracy when classifying patients with depression or bipolar disorder using a SVM classifier than using a Gaussian-process algorithm. Patel *et al.* (26) found that a decision tree algorithm (75%) outperformed linear SVMs (70%) and radial bias function-SVMs (68%). Interestingly, to date there are no systematic investigations of different algorithms in neuroimaging-based classification in psychiatry. It is noteworthy that choice of classification algorithm did not appear to affect performance in the classification of individuals diagnosed with schizophrenia (17). However, it needs to be noted that in the current meta-analysis all studies were of limited sample size so that potential differences between classification algorithms might not have manifested. Moreover, some algorithms require more extensive training samples and might not have been employed due to the limited amount of available training data. Another important factor in the context of our analysis is the embedding of feature selection, classifier optimization and the estimation of the models' generalizability in a CV scheme. We should note that in three papers that were included in the present meta-analysis, a feature selection procedure was implemented outside of the cross-validation (20,31,64). However, it is critical to avoid information leakage between the training and the test samples to avoid overfitting and biased estimates of classification accuracy. Moreover, even in the case of correct embedding, different cross-validation schemes might lead to different results. In our analysis, two-fold CV was associated with higher diagnostic accuracy than were 10-fold or leave-one-out CV.

Limitations of the Current Meta-analysis: Effect of Clinical Symptoms and Antidepressant Medication

One study suggested that the degree of functional and structural brain abnormalities found in depression is related to the severity of clinical symptomatology (76). In effect, neuroimaging-based classification models should perform better in more severely ill subjects. Mwangi *et al.* (76) reported a correlation between scores on the Beck Depression Inventory II and the Spielberger State and Trait Anxiety Inventory with individual SVM decision weights. In the present univariate analysis, we found no effect of clinical symptoms as measured by HAMD. It is possible that the clinical and methodological heterogeneity present in our meta-analysis obfuscated a potential relation between accuracy and clinical symptoms. On the other hand, using a multivariate meta-learning model, we found some evidence that depression severity as measured by the HAMD scale is an important predictor of classification accuracy. There are other clinical variables besides severity that may influence brain anomalies in patients with depression and that may affect the accuracy of neuroimaging-based classification. These variables include age of onset or illness duration and comorbidities such as anxiety, obsessive symptoms, or substance abuse. Unfortunately, few studies assessed in the current analysis reported this information. Thus effects of these variables could not be investigated in the present meta-analysis.

A potential confounding factor in the context of our analysis is antidepressant medication. Multiple studies reported changes in brain structure (77) and function (78–80) following chronic antidepressant treatment. If such effects are present in neuroimaging-based classification experiments, the brain patterns identified might be associated with drug effects rather than with effects specific to the pathology of depression. In our recent analysis in schizophrenia, we demonstrated that antipsychotic medication represents such potential bias (17). Because few studies in the present analysis reported treatment status, it was not possible to assess the potential impact of this factor.

Future Challenges for Neuroimaging-Based Classification of Depression

It is noteworthy that a substantial proportion of subjects included in the current meta-analysis (~25%) were not classified correctly. Multiple factors might drive misclassification using neuroimaging-based models. For example, the pattern

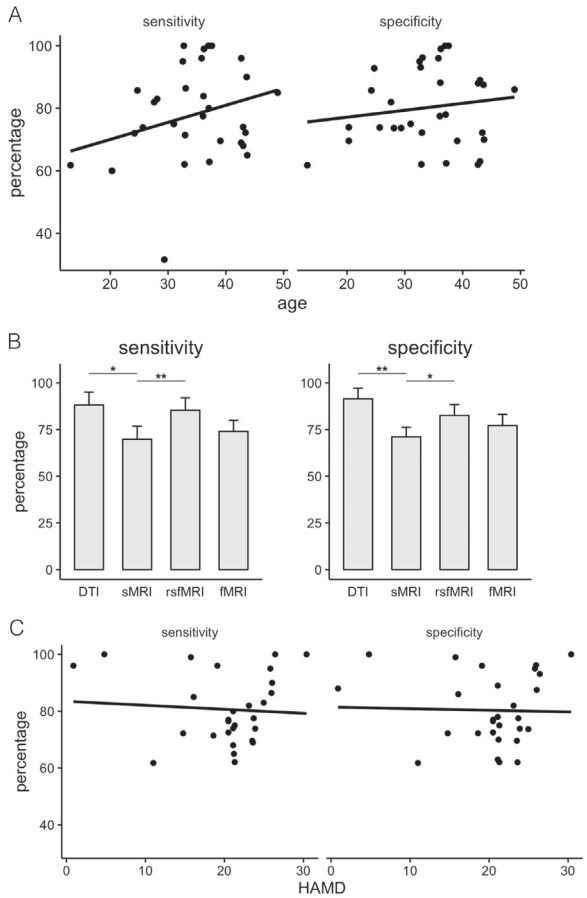

Figure 3. Results from the moderator analysis: **(A)** effect of age, **(B)** differences in sensitivity and specificity between imaging modalities, and **(C)** clinical symptoms as measured by HAMD. DTI, diffusion tensor imaging; fMRI, functional magnetic resonance imaging; HAMD, Hamilton Depression Rating Scale; rsfMRI, resting-state functional magnetic resonance imaging; sMRI, structural magnetic resonance imaging. $^{*}p < .05$; $^{**}p < .01$.

of brain changes associated with depression might have limited discriminative power. Alternatively, there might be one or more subgroups within individuals diagnosed with MDD that showed specific patterns of brain changes that are not shared by the majority of patients with depression. To explicitly test this hypothesis, a more detailed investigation of potential moderator variables on the basis of the data of individual subjects is required. Moreover, future studies that focus on unsupervised classification methods would be better suited to identify such subgroups. This approach to classification might account for heterogeneity and improve diagnostic accuracy. Finally, the performance of neuroimaging-based diagnostic models is limited by the reliability of the diagnostic labels. For example, the recent investigation of the test-retest reliability of the diagnostic categories of the DSM-5 indicated that MDD was diagnosed with limited accuracy (81). Similarly, MDD and bipolar disorder might frequently be confused (82). In effect, misclassification in the initial psychiatric assessment might have contributed to reductions in classification performance.

An important consideration in the context of neuroimaging-based disease classification is the differentiation between diagnoses. Rather than distinguishing patients from HCs, a substantial part of clinical practice involves laborious and error-prone differential diagnostic processes to distinguish different patient groups from each other rather than from HCs. To date, few studies have investigated the potential of neuroimaging-based diagnostic models to differentiate among diagnostic groups. For example, Redlich *et al*. (19) studied two independent samples in which they were able to differentiate depressed from bipolar patients with 79.3% and 65.5% accuracy based on sMRI data, whereas Sacchet *et al*. (20) reported a lower classification accuracy of 59.5%. Similarly, in our recent work, we could demonstrate that patients with schizophrenia can be separated from depressed patients with 76% accuracy based on sMRI data (18). However, Serpa *et al*. (21) reported only 54% accuracy when differentiating psychotic bipolar patients from psychotic depressed patients, suggesting that these patient groups are harder to separate using brain-based features. In summary, differential-diagnostics represent an interesting potential application of neuroimaging-based models in clinical practice and a way to validate the current diagnostic categories in psychiatry.

Another important challenge for the application of neuroimaging-based diagnostics involves the generalizability of diagnostic models across different sites and populations. Redlich *et al.* (19) demonstrated that neuroimaging-based classifiers can be trained on data acquired at one site and then be applied to data from a different site. Similarly, Koutsouleris *et al.* (83) have also demonstrated the feasibility of such cross-site neuroimaging-based classification for patients with schizophrenia. However, it needs to be noted that all studies included in the present meta-analysis were of small or modest sample size and that recent analyses suggest that larger samples are required for reliable estimates (84). A large-scale investigation of the generalizability of neuroimaging-based models (e.g., in the form of an individual patient data meta-analysis or mega-analysis) is still missing and the clinical and methodological factors that influence generalization are not clear.

We should note that whereas the present analyses support the hypothesis that multivariate methods are able to identify biological signatures of MDD in neuroimaging data, the current accuracy of ~75% does not allow direct clinical application of these models. Moreover, neuroimaging-based diagnostic models must be evaluated critically with respect to cost efficiency. For example, self-rated screening questionnaires such as the nine-item Patient Health Questionnaire provide an estimated sensitivity of 77% and a specificity of 85% in identifying MDD (85). Moreover, clinical questionnaires can be administered at substantially lower costs compared with neuroimaging investigations. Therefore, the main potential of neuroimaging-based diagnostic models might be to predict response to treatment interventions or to predict the course of the disorder. Generally subjects receiving pharmacological or psychotherapeutic interventions show large heterogeneity with respect to improvement or side effects (86). Patel *et al.* (26) showed that response to treatment with a selective serotonin reuptake inhibitor could be predicted with up to 89% accuracy using neuroimaging. Relatedly, Fu *et al*. (59) used brain

activation from a fMRI scan before treatment initiation to predict partial response to antidepressant treatment with 75% and full response with 62% accuracy. In another study, Siegle *et al.* (87) found that brain activation in response to negative words predicts response to cognitive-behavioral therapy with 75% and remission with 70% of accuracy. Using DTI data, Korgaonkar *et al.* (66) reported the prediction of treatment response in major depression with 74% accuracy. Relatedly, Lythe *et al.* (88) reported that neuroimaging-based models allow the prediction of recurrence risk of medication-free patients with MDD with 75% accuracy. In summary, neuroimaging-based classification represents a promising approach for classification of subjects with depression. Moreover, this approach might be of benefit to other endeavors, such as the prediction of disease course or treatment outcome. Current limitations include the generalizability of the models across research centers and the identification of methodological and clinical variables that moderate classification success.

ACKNOWLEDGMENTS AND DISCLOSURES

JK is supported by funds from the Friedrich-Baur Stiftung as well as the Förderung Forschung und Lehre (881/856).

We would like to thank the authors of the included studies for providing additional information.

All authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Psychiatry (JK, CC, PF, NK), Ludwig-Maximilians University Munich, Munich; Clinical Affective Neuroimaging Laboratory (MW), Department of Behavioural Neurology, Leibniz Institute for Neurobiology, Magdeburg; Department of Psychiatry and Psychotherapy (MW), Eberhard Karls University, Tubingen, Germany; Neurosciences Program and Department of Psychology (MDS, IHG), Stanford University, Stanford, California; Institute of Psychiatry (RZ), King's College London, London, United Kingdom; Laboratory of Psychiatric Neuroimaging (MHS), Institute and Department of Psychiatry; and Center for Interdisciplinary Research on Applied Neurosciences (NAPNA) (MHS), University of Sao Paulo, Sao Paulo, Brazil.

Address correspondence to Joseph Kambeitz M.D., Dipl.-Psych., Nußbaumstraße 7, 80336 Munich, Germany; E-mail: joseph.kambeitz@med.uni-muenchen.de.

Received May 1, 2016; revised Sep 27, 2016; accepted Oct 20, 2016.

Supplementary material cited in this article is available online at http://dx.doi.org/10.1016/j.biopsych.2016.10.028.

REFERENCES

1. Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, de Girolamo G, *et al*. (2011): Cross-national epidemiology of DSM-IV major depressive episode. BMC Med 9:90.
2. Albert PR, Benkelfat C (2013): The neurobiology of depression—revisiting the serotonin hypothesis. II. Genetic, epigenetic and clinical studies. Philos Trans R Soc Lond B Biol Sci 368:20120535.
3. Sacher J, Neumann J, Fünfstück T, Soliman A, Villringer A, Schroeter ML (2012): Mapping the depressed brain: A meta-analysis of structural and functional alterations in major depressive disorder. J Affect Disord 140:142–148.
4. Sexton CE, Mackay CE, Ebmeier KP (2009): A systematic review of diffusion tensor imaging studies in affective disorders. Biol Psychiatry 66:814–823.
5. Schmaal L, Veltman DJ, van Erp TGM, Sämann PG, Frodl T, Jahanshad N, *et al*. (2016): Subcortical brain alterations in major depressive disorder: Findings from the ENIGMA Major Depressive Disorder working group. Mol Psychiatry 21:806–812.
6. Diener C, Kuehner C, Brusniak W, Ubl B, Wessa M, Flor H (2012): A meta-analysis of neurofunctional imaging studies of emotion and cognition in major depression. Neuroimage 61:677–685.
7. Kaiser RH, Andrews-Hanna JR, Wager TD, Pizzagalli DA (2015): Large-scale network dysfunction in major depressive disorder: A meta-analysis of resting-state functional connectivity. JAMA Psychiatry 72:603–611.
8. Kühn S, Gallinat J (2011): Resting-state brain activity in schizophrenia and major depression: A quantitative meta-analysis. Schizophr Bull 39:358–365.
9. Gryglewski G, Lanzenberger R, Kranz GS, Cumming P (2014): Meta-analysis of molecular imaging of serotonin transporters in major depression. J Cereb Blood Flow Metab 34:1096–1103.
10. Kambeitz JP, Howes OD (2015): The serotonin transporter in depression: Meta-analysis of in vivo and post mortem findings and implications for understanding and treating depression. J Affect Disord 186:358–366.
11. Fried EI, Kievit RA (2016): The volumes of subcortical regions in depressed and healthy individuals are strikingly similar: A reinterpretation of the results by Schmaal *et al*. Mol Psychiatry 21:724–725.
12. Borgwardt S, Radua J, Mechelli A, Fusar-Poli P (2012): Why are psychiatric imaging methods clinically unreliable? Conclusions and practical guidelines for authors, editors and reviewers. Behav Brain Funct 8:46.
13. Kapur S, Phillips AG, Insel TR (2012): Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol Psychiatry 17:1174–1179.
14. Davatzikos C (2004): Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. Neuroimage 23:17–20.
15. Klöppel S, Abdulkadir A, Jack CR Jr, Koutsouleris N, Mourão-Miranda J, Vemuri P (2012): Diagnostic neuroimaging across diseases. Neuroimage 61:457–463.
16. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF (2015): From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev 57:328–349.
17. Kambeitz J, Kambeitz-Ilankovic L, Leucht S, Wood S, Davatzikos C, Malchow B, *et al*. (2015): Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. Neuropsychopharmacology 40:1742–1751.
18. Koutsouleris N, Meisenzahl EM, Borgwardt S, Riecher-Rössler A, Frodl T, Kambeitz J, *et al*. (2015): Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. Brain 138:2059–2073.
19. Redlich R, Almeida JJ, Grotegerd D, Opel N, Kugel H, Heindel W, *et al*. (2014): Brain morphometric biomarkers distinguishing unipolar and bipolar depression: A voxel-based morphometry-pattern classification approach. JAMA Psychiatry 71:1222–1230.
20. Sacchet MD, Livermore EE, Iglesias JE, Glover GH, Gotlib IH (2015): Subcortical volumes differentiate major depressive disorder, bipolar disorder, and remitted major depressive disorder. J Psychiatr Res 68:91–98.
21. Serpa MH, Ou Y, Schaufelberger MS, Doshi J, Ferreira LK, Machado-Vieira R, *et al*. (2014): Neuroanatomical classification in a population-based sample of psychotic major depression and bipolar I disorder with 1 year of diagnostic stability. Biomed Res Int 2014:706157.
22. Kambeitz-Ilankovic L, Meisenzahl EM, Cabral C, von Saldern S, Kambeitz J, Falkai P, *et al*. (2016): Prediction of outcome in the psychosis prodrome using neuroanatomical pattern classification. Schizophr Res 173:159–165.
23. Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, *et al*. (2009): Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch Gen Psychiatry 66:700–712.
24. Mourao-Miranda J, Reinders AA, Rocha-Rego V, Lappin J, Rondina J, Morgan C, *et al*. (2012): Individualized prediction of illness course at

the first psychotic episode: A support vector machine MRI study. Psychol Med 42:1037–1047.
25. Mwangi B, Ebmeier KP, Matthews K, Steele JD (2012): Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain 135: 1508–1521.
26. Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF 3rd, Aizenstein HJ (2015): Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int J Geriatr Psychiatry 30:1056–1067.
27. Wu MJ, Wu HE, Mwangi B, Sanches M, Selvaraj S, Zunta-Soares GB, Soares JC (2015): Prediction of pediatric unipolar depression using multiple neuromorphometric measurements: A pattern classification approach. J Psychiatr Res 62:84–91.
28. Cao L, Guo S, Xue Z, Hu Y, Liu H, Mwansisya TE, *et al*. (2014): Aberrant functional connectivity for diagnosis of major depressive disorder: A discriminant analysis. Psychiatry Clin Neurosci 68: 110–119.
29. Craddock RC, Holtzheimer PE, Hu XP, Mayberg HS (2009): Disease state prediction from resting state functional connectivity. Magn Reson Med 62:1619–1628.
30. Guo H, Cheng C, Cao X, Xiang J, Chen J, Zhang K (2014): Resting-state functional connectivity abnormalities in first-onset unmedicated depression. Neural Regen Res 9:153–163.
31. Lord A, Horn D, Breakspear M, Walter M (2012): Changes in community structure of resting state functional connectivity in unipolar depression. PLoS One 7:e41282.
32. Ma Q, Zeng LL, Shen H, Liu L, Hu D (2013): Altered cerebellar-cerebral resting-state functional connectivity reliably identifies major depressive disorder. Brain Res 1495:86–94.
33. Wei M, Qin J, Yan R, Li H, Yao Z, Lu Q (2013): Identifying major depressive disorder using Hurst exponent of resting-state brain networks. Psychiatry Res 214:306–312.
34. Zeng LL, Shen H, Liu L, Wang L, Li B, Fang P, *et al*. (2012): Identifying major depression using whole-brain functional connectivity: A multivariate pattern analysis. Brain 135:1498–1507.
35. Habes I, Krall SC, Johnston SJ, Yuen KS, Healy D, Goebel R, *et al*. (2013): Pattern classification of valence in depression. Neuroimage Clin 2:675–683.
36. Hahn T, Marquand AF, Ehlis AC, Dresler T, Kittel-Schneider S, Jarczok TA, *et al*. (2011): Integrating neurobiological markers of depression. Arch Gen Psychiatry 68:361–368.
37. Marquand AF, Mourão-Miranda J, Brammer MJ, Cleare AJ, Fu CH (2008): Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. Neuroreport 19:1507–1511.
38. Nouretdinov I, Costafreda SG, Gammerman A, Chervonenkis A, Vovk V, Vapnik V, Fu CH (2011): Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. Neuroimage 56:809–813.
39. Rosa MJ, Portugal L, Hahn T, Fallgatter AJ, Garrido MI, Shawe-Taylor J, Mourao-Miranda J (2015): Sparse network-based models for patient classification using fMRI. Neuroimage 105:493–506.
40. Moher D, Liberati A, Tetzlaff J, Altman DG, for the PRISMA Group (2009): Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ 339:b2535.
41. Hamilton M (1960): A rating scale for depression. J Neurol Neurosurg Psychiatry 23:56–62.
42. Beck AT, Steer RA, Ball R, Ranieri W (1996): Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. J Pers Assess 67:588–597.
43. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH (2005): Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 58:982–990.
44. Deeks JJ, Macaskill P, Irwig L (2005): The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 58:882–893.
45. R Core Team (2013): R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org/. Accessed January 1, 2016.
46. Doebler P (2012): Meta-analysis of diagnostic accuracy with mada. Available at: http://cran.gis-lab.info/web/packages/mada/vignettes/mada.pdf. Accessed May 18, 2013.
47. Friedman J, Hastie T, Tibshirani R (2010): Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22.
48. Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, *et al*. (2011): Prognostic prediction of therapeutic response in depression using high-field MR imaging. Neuroimage 55:1497–1503.
49. Liu F, Guo W, Yu D, Gao Q, Gao K, Xue Z, *et al*. (2012): Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS One 7:e40968.
50. Kipli K, Kouzani AZ (2014): Degree of contribution (DoC) feature selection algorithm for structural brain MRI volumetric features in depression detection. Int J Comput Assist Radiol Surg 10: 1003–1016.
51. Rondina JM, Hahn T, de Oliveira L, Marquand AF, Dresler T, Leitner T, *et al*. (2014): SCoRS—A method based on stability for feature selection and mapping in neuroimaging [corrected]. IEEE Trans Med Imaging 33:85–98.
52. Johnston BA, Steele JD, Tolomeo S, Christmas D, Matthews K (2015): Structural MRI-based predictions in patients with treatment-refractory depression (TRD). PLoS One 10:e0132958.
53. Wade BS, Joshi SH, Pirnia T, Leaver AM, Woods RP, Thompson PM, *et al*. (2015): Random forest classification of depression status based on subcortical brain morphometry following electroconvulsive therapy. Proc IEEE Int Symp Biomed Imaging 2015:92–96.
54. Jie NF, Zhu MH, Ma XY, Osuch EA, Wammes M, Théberge J, *et al*. (2015): Discriminating bipolar disorder from major depression based on SVM-FoBa: Efficient feature selection with multimodal brain imaging data. IEEE Trans Auton Ment Dev 7:320–331.
55. Guo H, Cao X, Liu Z, Li H, Chen J, Zhang K (2012): Machine learning classifier using abnormal brain network topological metrics in major depressive disorder. Neuroreport 23:1006–1011.
56. Guo S, Yu Y, Zhang J, Feng J (2013): A reversal coarse-grained analysis with application to an altered functional circuit in depression. Brain Behav 3:637–648.
57. Lois G, Wessa M (2016): Differential association of default mode network connectivity and rumination in healthy individuals and remitted MDD patients. Soc Cogn Affect Neurosci 11:1792–1801.
58. Qin J, Shen H, Zeng LL, Jiang W, Liu L, Hu D (2015): Predicting clinical responses in major depression using intrinsic functional connectivity. Neuroreport 26:675–680.
59. Fu CH, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SC, Brammer MJ (2008): Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression. Biol Psychiatry 63:656–662.
60. Zhang X, Yaseen ZS, Galynker II, Hirsch J, Winston A (2011): Can depression be diagnosed by response to mother's face? A personalized attachment-based paradigm for diagnostic fMRI. PLoS One 6: e27253.
61. Yang W, Chen Q, Liu P, Cheng H, Cui Q, Wei D, *et al*. (2016): Abnormal brain activation during directed forgetting of negative memory in depressed patients. J Affect Disord 190:880–888.
62. Sato JR, Moll J, Green S, Deakin JF, Thomaz CE, Zahn R (2015): Machine learning algorithm accurately detects fMRI signature of vulnerability to major depression. Psychiatry Res 233:289–291.
63. Shimizu Y, Yoshimoto J, Toki S, Takamura M, Yoshimura S, Okamoto Y, *et al*. (2015): Toward probabilistic diagnosis and understanding of depression based on functional MRI data analysis with logistic group LASSO. PLoS One 10:e0123524.
64. Qin J, Wei M, Liu H, Chen J, Yan R, Yao Z, Lu Q (2015): Altered anatomical patterns of depression in relation to antidepressant treatment: Evidence from a pattern recognition analysis on the topological organization of brain networks. J Affect Disord 180: 129–137.

65. Fang P, Zeng LL, Shen H, Wang L, Li B, Liu L, Hu D (2012): Increased cortical-limbic anatomical network connectivity in major depression revealed by diffusion tensor imaging. PLoS One 7:e45972.
66. Korgaonkar MS, Williams LM, Song YJ, Usherwood T, Grieve SM (2014): Diffusion tensor imaging predictors of treatment outcomes in major depressive disorder. Br J Psychiatry 205:321–328.
67. Sacchet MD, Prasad G, Foland-Ross LC, Thompson PM, Gotlib IH (2015): Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. Front Psychiatry 6:21.
68. Nugent AC, Neumeister A, Goldman D, Herscovitch P, Charney DS, Drevets WC (2008): Serotonin transporter genotype and depressive phenotype determination by discriminant analysis of glucose metabolism under acute tryptophan depletion. Neuroimage 43:764–774.
69. Van Dijk KR, Sabuncu MR, Buckner RL (2012): The influence of head motion on intrinsic functional connectivity MRI. Neuroimage 59:431–438.
70. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012): Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59:2142–2154.
71. Satterthwaite TD, Wolf DH, Loughead J, Ruparel K, Elliott MA, Hakonarson H, *et al*. (2012): Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. Neuroimage 60:623–632.
72. Cao H, Plichta MM, Schäfer A, Haddad L, Grimm O, Schneider M, *et al*. (2014): Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. Neuroimage 84:888–900.
73. Eaton WW, Anthony JC, Gallo J, Cai G, Tien A, Romanoski A, *et al*. (1997): Natural history of Diagnostic Interview Schedule/DSM-IV major depression: The Baltimore Epidemiologic Catchment Area follow-up. Arch Gen Psychiatry 54:993–999.
74. Cabral C, Kambeitz-Ilankovic L, Kambeitz J, Calhoun VD, Dwyer DB, von Saldern S, *et al*. (2016): Classifying schizophrenia using multimodal multivariate pattern recognition analysis: evaluating the impact of individual clinical profiles on the neurodiagnostic performance. Schizophr Bull 42(suppl 1):S110–S117.
75. Sui J, Castro E, He H, Bridwell D, Du Y, Pearlson GD, *et al*. (2014): Combination of FMRI-SMRI-EEG data improves discrimination of schizophrenia patients by ensemble feature selection. Conf Proc IEEE Eng Med Biol Soc 2014:3889–3892.
76. Mwangi B, Matthews K, Steele JD (2012): Prediction of illness severity in patients with major depression using structural MR brain scans. J Magn Reson Imaging 35:64–71.
77. Jung J, Kang J, Won E, Nam K, Lee MS, Tae WS, Ham BJ (2014): Impact of lingual gyrus volume on antidepressant response and neurocognitive functions in Major Depressive Disorder: A voxel-based morphometry study. J Affect Disord 169:179–187.
78. Abler B, Grön G, Hartmann A, Metzger C, Walter M (2012): Modulation of frontostriatal interaction aligns with reduced primary reward processing under serotonergic drugs. J Neurosci 32:1329–1335.
79. Metzger CD, Wiegers M, Walter M, Abler B, Graf H (2015): Local and global resting state activity in the noradrenergic and dopaminergic pathway modulated by reboxetine and amisulpride in healthy subjects. Int J Neuropsychopharmacol 19:pii:pyv080.
80. Posner J, Hellerstein DJ, Gat I, Mechling A, Klahr K, Wang Z, *et al*. (2013): Antidepressants normalize the default mode network in patients with dysthymia. JAMA Psychiatry 70:373–382.
81. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ (2013): DSM-5 field trials in the United States and Canada, part II: Test-retest reliability of selected categorical diagnoses. Am J Psychiatry 170:59–70.
82. Oiesvold T, Nivison M, Hansen V, Skre I, Ostensen L, Sørgaard KW (2013): Diagnosing comorbidity in psychiatric hospital: Challenging the validity of administrative registers. BMC Psychiatry 13:13.
83. Koutsouleris N, Davatzikos C, Borgwardt S, Gaser C, Bottlender R, Frodl T, *et al*. (2014): Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. Schizophr Bull 40:1140–1153.
84. Schnack HG, Kahn RS (2016): Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. Front Psychiatry 7:50.
85. Manea L, Gilbody S, McMillan D (2015): A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen Hosp Psychiatry 37:67–75.
86. Cipriani A, Purgato M, Furukawa TA, Trespidi C, Imperadore G, Signoretti A, *et al*. (2012): Citalopram versus other anti-depressive agents for depression. Cochrane Database Syst Rev 7: CD006534.
87. Siegle GJ, Thompson WK, Collier A, Berman SR, Feldmiller J, Thase ME, Friedman ES (2012): Toward clinically useful neuroimaging in depression treatment: Prognostic utility of subgenual cingulate activity for determining depression outcome in cognitive therapy across studies, scanners, and patient characteristics. Arch Gen Psychiatry 69:913–924.
88. Lythe KE, Moll J, Gethin JA, Workman CI, Green S, Lambon Ralph MA, *et al*. (2015): Self-blame-selective hyperconnectivity between anterior temporal and subgenual cortices and prediction of recurrent depressive episodes. JAMA Psychiatry 72:1119–1126.
89. Zwinderman A, Bossuyt P (2008): We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med 27:687–697.