

ARTICLE

Open Access

Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: a systematic review and meta-analysis

Sem E. Cohen¹, Jasper B. Zantvoord ^{1,2}, Babet N. Wezenberg¹, Claudi L. H. Bockting^{1,3} and Guido A. van Wingen¹

Abstract

No tools are currently available to predict whether a patient suffering from major depressive disorder (MDD) will respond to a certain treatment. Machine learning analysis of magnetic resonance imaging (MRI) data has shown potential in predicting response for individual patients, which may enable personalized treatment decisions and increase treatment efficacy. Here, we evaluated the accuracy of MRI-guided response prediction in MDD. We conducted a systematic review and meta-analysis of all studies using MRI to predict single-subject response to antidepressant treatment in patients with MDD. Classification performance was calculated using a bivariate model and expressed as area under the curve, sensitivity, and specificity. In addition, we analyzed differences in classification performance between different interventions and MRI modalities. Meta-analysis of 22 samples including 957 patients showed an overall area under the bivariate summary receiver operating curve of 0.84 (95% CI 0.81–0.87), sensitivity of 77% (95% CI 71–82), and specificity of 79% (95% CI 73–84). Although classification performance was higher for electroconvulsive therapy outcome prediction ($n = 285$, 80% sensitivity, 83% specificity) than medication outcome prediction ($n = 283$, 75% sensitivity, 72% specificity), there was no significant difference in classification performance between treatments or MRI modalities. Prediction of treatment response using machine learning analysis of MRI data is promising but should not yet be implemented into clinical practice. Future studies with more generalizable samples and external validation are needed to establish the potential of MRI to realize individualized patient care in MDD.

Introduction

Major depressive disorder (MDD) is a debilitating disease, accounting for 40% of the global disability-adjusted life years caused by psychiatric disorders¹. Depression is associated with impaired social functioning and unemployment and is associated with a wide range of chronic physical illnesses, such as diabetes and cardiovascular disease^{2,3}. MDD is estimated to have a life-time

prevalence of 20.6% in the United States⁴. Despite general consensus that effective treatment of depression is paramount for both a patient's health and for reducing global burden of disease, global disease burden by MDD has not decreased in the past decades⁵. This is partly because treatment selection is based on trial and error, with no possibility to predict an individual's response to a certain treatment⁶. Non-response to initial pharmacological and psychotherapeutic interventions is highly prevalent, with treatment-resistant depression affecting 20–30% of depressed patients in the current clinical practice^{7–9}. Treatment of choice for patients who have not responded to pharmacological and psychotherapeutic treatments is electroconvulsive therapy (ECT), which

Correspondence: Guido A. van Wingen (g.a.vanwingen@amsterdamumc.nl)

¹Department of Psychiatry, Amsterdam Neuroscience, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

²Department of Child and Adolescent Psychiatry, Amsterdam University Medical Center, Amsterdam, The Netherlands

Full list of author information is available at the end of the article

© The Author(s) 2021



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

produces remission in about 50% of therapy-resistant patients^{10,11}. Furthermore, non-response can only be determined at least 4 weeks after initiation of pharmacotherapy, ECT requires 4–6 weeks on average, and effects of psychotherapy can even take 16 weeks to manifest^{7,12}. Consequently, patients are regularly exposed to multiple failed treatments and might spend months to years waiting for successful treatment. This stresses the need for markers, which, before treatment commencement, can inform clinicians on the chance of responding to a particular treatment.

A large number of studies have correlated baseline clinical characteristics and biomarkers with MDD status and treatment outcome and have identified many factors that are associated with treatment success¹³. However, such descriptive analyses only provide inference at the group level and not at the level of the individual patient, which is required for clinical decision-making¹⁴. More recent studies have started to use machine learning analyses that aim to develop predictive models and that are tested using independent data¹⁵. More than with correlational analysis, single-subject response prediction studies using machine learning might be able to redeem the promise of individualized psychiatry¹⁶. Without being explicitly pre-programmed, these algorithms (either linear or non-linear) are able to learn from aggregated data in a patient sample using multivariate pattern recognition, in order to provide the best prediction of an output variable^{17,18}. In predictive modeling, machine learning could enable clinicians to judge the viability of treatments for individual patients. As such, it might increase treatment efficacy, decrease illness duration, and reduce MDD's impact on the global burden of disease.

Multiple modalities have been considered for single-subject response prediction. A recent meta-analysis covering different markers found neuroimaging to overall be most successful in predicting treatment response in depressed patients (i.e., more than phenomenological or genetic studies)¹⁹. However, the review pooled different treatments and neuroimaging modalities such as electroencephalography (EEG) and magnetic resonance imaging (MRI). Since it did not differentiate between prediction success in different neuroimaging techniques, the study offers little insight into treatment-specific biomarkers or specific (MRI) modalities. A recent meta-analysis on EEG for individual prediction of antidepressant treatment response found reasonable accuracy (72% sensitivity and 68% specificity) but concludes that EEG should not yet be used clinically as a prediction tool, since generalizability and validity of the reported studies are limited²⁰. However, a meta-analysis of prediction accuracy in antidepressant treatment that specifically focuses on MRI does not yet exist, which may reveal a better predictive value than EEG.

The primary aim of the present study was to calculate the aggregate classification performance of predictive MRI biomarkers in patients with MDD using a bivariate random-effect model meta-analysis. We further investigated whether classification performance was influenced by intervention type (i.e., pharmacotherapy, psychotherapy, or ECT) or imaging modality (i.e., structural MRI (sMRI), resting-state functional MRI (fMRI), task-based fMRI, diffusion tensor imaging (DTI)).

Methods and materials

Inclusion and exclusion criteria

Two authors (S.E.C. and B.N.W.) included studies using any form of MRI (structural, resting-state, task-based, spectroscopy, DTI), which were conducted at baseline, i.e., within 4 weeks before the start of antidepressant treatment. Furthermore, inclusion criteria were an overarching definition of antidepressant treatment according to the current NICE guidelines and a non-selective patient population with MDD suffering from a current depressive episode. Studies that used feature selection based on in-sample data without validating prediction outcomes either internally (e.g., through cross-validation) or externally (through independent set validation) were excluded. Inclusion or exclusion conflicts were resolved by consensus or if necessary by authors J.B.Z. and G.A.v.W.

Search strategy

We conducted a search in EMBASE, Medline, PsycInfo, and Web of Science databases. Each database was searched from inception to January 2020. Furthermore, we searched the WHO International Clinical Trial Registry Platforms search portal for registered and unpublished studies, and we looked for “gray” literature such as abstracts and conference articles through conference websites and from other relevant sources. Additionally, we checked included articles for references and conducted citation screening. For a full account of our search strategy and inclusion criteria, see the Supplementary Material.

Data extraction

Two authors (S.E.C. and B.N.W.) independently extracted data from included studies, including the number of participants, patient population and depression severity subtype, treatment history, antidepressant intervention and outcome measures, response/remission rates, neuroimaging technique, brain region and feature selection, method of analysis, and validation strategy (see Table 1). From the included articles, we extracted the confusion table (a 2 × 2 table for correctly and incorrectly classified patients) for sensitivity or specificity. If these were not supplied, we computed the matrix from additional information in the article. If multiple studies analyzed the same

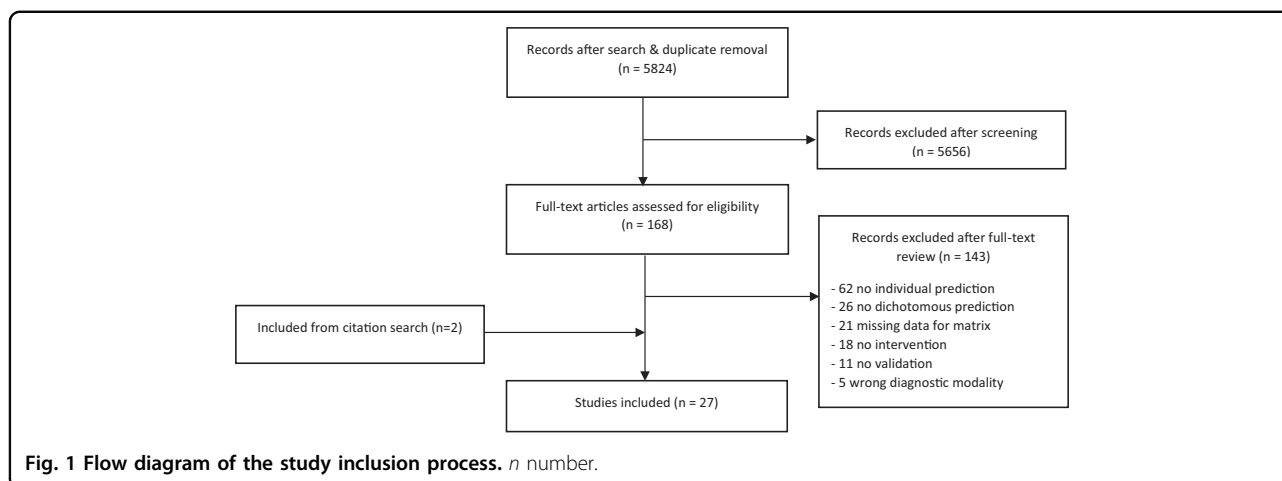
Table 1 Methodological summary of the studies.

Study + year	n	Outcome	Intervention	Duration	Modality	Analysis	Validation
Costafreda et al. 2009—1	16	Remission	CBT	16 wk	tbfMRI	SVM	LOO CV
Siegle et al. 2012	12	Response	CBT	12 wk	tbfMRI	RF	Ind. replication
Queirazza et al. 2019	37	Response	CBT	6–10 wk	tbfMRI	SVM, LR	LOO nested CV
Van Waarde et al. 2015	45	Remission	ECT	10 wk	rsfMRI	ISVM	LOO CV
Moreno-Ortega et al. 2019	19	Remission	ECT	ns	rsfMRI	LR	LOO CV
Sun et al. 2019	122	Remission + remission	ECT	3–4 wk	rsfMRI	LR	LOO CV
Redlich et al. 2016	23	Response	ECT	3–8 wk	sMRI	ISVM/GPC	LOO CV
Wade et al. 2016	34	Response	ECT	2–7 wk	sMRI	RBFSVM	LOO CV
Cao et al. 2018	24	Response + remission	ECT	3–4 wk	sMRI	ISVM	LOO CV
Jiang et al. 2018	38	Remission	ECT	3–4 wk	sMRI	LR	10-fold LOO CV+ Independent cohort rep
Wade et al. 2017	44 ^a	Remission	ECT	ns	sMRI	RF	Nested CV
Leaver et al. 2017		Response	ECT	ns	rsfMRI, aslMRI	RBFSVM	5-fold LOO CV
Drysdale et al. 2017	124 30	Response	rTMS	4–6 wk	rsfMRI	ISVM	LOO CV Ind. replication
Cash et al. 2019	33	Remission	rTMS	5–8 wk	rsfMRI	ISVM	LOO + k fold CV
Costafreda et al. 2009—2	18 ^a	Remission	SSRI	8 wk	sMRI	ISVM	LOO CV
Nouretdinov et al. 2011		Remission	SSRI	8 wk	sMRI	TCP	LOO CV
Gong et al. 2011	46	Response	SSRI/TCA/SNRI	12 wk	sMRI	ISVM	LOO CV
Marquand et al. 2008	20	Response	SSRI	8 wk	tbfMRI	ISVM	LOO CV
Godlewska et al. 2018	32	Response	SSRI	6 wk	tbfMRI	LR	LOO CV
Meyer et al. 2019	22	Remission/non-response	SSRI	8 wk	tbfMRI	LR	LOO CV
Karim et al. 2018	49	Remission	SNRI	12 wk	tbfMRI	LR	10-fold LOO CV
Patel et al. 2015	19	Remission	SSRI/SNRI	ns	rsfMRI, DTI, sMRI	ADTree/ISVM/ RBFSVM/L1LR	Nested LOO CV
<i>iSPOT trials</i>	77 ^a		SSRI/SNRI	8 wk			
Korgaonkar et al. 2014		Remission			DTI	LR	K-fold CV
Williams et al. 2015		Response			tbfMRI	LDA	LOO CV
Goldstein-Piekarski et al. 2016		Remission			tbfMRI	LR	10-fold LOO CV
Grieve et al. 2016		Non-remission			DTI	LR	Independent rep
Goldstein-Piekarski et al. 2018		Remission			rsfMRI	LR	LOO CV

Reported sample sizes were not necessarily equal in articles with overlapping sample.

SSRI selective serotonin reuptake inhibitor, TCA tricyclic antidepressant, SNRI serotonin-norepinephrine reuptake inhibitor, ECT electroconvulsive therapy, CBT cognitive behavioral therapy, rTMS repetitive transcranial magnetic stimulation, iTBS intermittent theta burst stimulation, AP antipsychotics, ns not specified, tb task based, rs resting state, asl arterial spin labeling, fMRI functional magnetic resonance imaging, sMRI structural magnetic resonance imaging, WB whole brain, ROI region of interest, DTI diffusion tensor imaging, ISVM linear support vector machine, RBF radial basic function, TCP transductive conformal predictor, LR logistic regression, LinR linear regression, LDA linear discriminant analysis, RF random forest, LOO CV leave-one-out cross-validation, wm white matter, sLR stepwise linear regression, beta-w beta-weights, LARS least-angle regression, PMVD proportional marginal decomposition.

^an is a weighted average across studies.



patient sample, we used mean outcome measures based on these studies. If necessary, we contacted authors requesting additional information.

Meta-analytic method

For quantitative analysis, we used confusion matrices to pool studies using Reitsma's bivariate random effect model, as suggested in the Cochrane handbook for diagnostic tests of accuracy studies^{21,22}. We used this method for computing our main outcomes, which were the overall area under the summary receiver operating characteristic (SROC) curve, sensitivity, and specificity, as well as sensitivity and specificity of intervention subsets. Additionally, we performed a separate bivariate regression for modalities (fMRI and sMRI) by including from each study both sMRI and fMRI, if provided in the original article or after our request for further information. As a post hoc analysis, we excluded DTI from this regression, and in the fMRI group, we subdivided resting-state and task-based modalities.

Heterogeneity and publication bias

To visualize between-study differences, we conducted a univariate random-effect forest plot of the diagnostic odds ratios (ORs), subdivided per treatment group. We identified clinical and statistical heterogeneity by visually assessing confidence interval (CI) overlap and by identifying outlying studies. We avoided using an objective measure of heterogeneity, since these have shown to be inappropriately conservative for accuracy studies²³. Rather, we used a random-effect model that assumes that our data was heterogeneous and set out to investigate potential sources of heterogeneity²². We did not perform any sensitivity analyses, as no studies were of such low quality, or were such outliers that sensitivity analysis was appropriate. To assess sample size effects and possible publication bias, we used Deeks' test, as recommended for diagnostic accuracy studies^{24,25}. For assessing quality of the primary studies, we

used the QUADAS-2 tool²⁶. We pre-specified methods in the PROSPERO database for systematic reviews (registration number CRD42019137497). All analyses were conducted using the mada and metafor package in R^{27–29}.

Results

Search results

Our search yielded 5824 hits, 168 of which were included for full-text review (see Fig. 1). After contacting the authors for additional information, we excluded 21 studies for not reporting data necessary for reconstructing a confusion matrix, all of which were “gray literature”, i.e., abstracts or conference summary articles. Furthermore, we excluded 11 articles for not reporting any form of validation of their prediction model. After exclusion of non-eligible studies and, through citation searching, addition of 2 eligible studies that did not come up in search hits, 27 remained^{30–56}.

Description of the study characteristics

We included 27 studies with an accumulated number of 957 unique patients and a mean sample size of 44 per study, with a median of 33 (see Table 1 for a full methodological study summary. Please refer to Supplementary Table 1 for an overview of patient characteristics and study demographics). Three patient samples were used in more than one article^{30,32,40,41,51–55}.

Of the included studies, 50% used some form of pharmacotherapeutic intervention (total $n = 283$), all of which administered a clinically viable dosage, with response time varying from 2 weeks (early response) to 12 weeks. Only one study did not use selective serotonin reuptake inhibitors (SSRIs), instead using an serotonin-norepinephrine reuptake inhibitor (SNRI)⁴⁹. Three studies used either an SSRI or SNRI, and one of these three chose a tricyclic antidepressant as a third treatment option^{45,50,57}. ECT was administered in 35% of studies (total $n = 285$), 8%

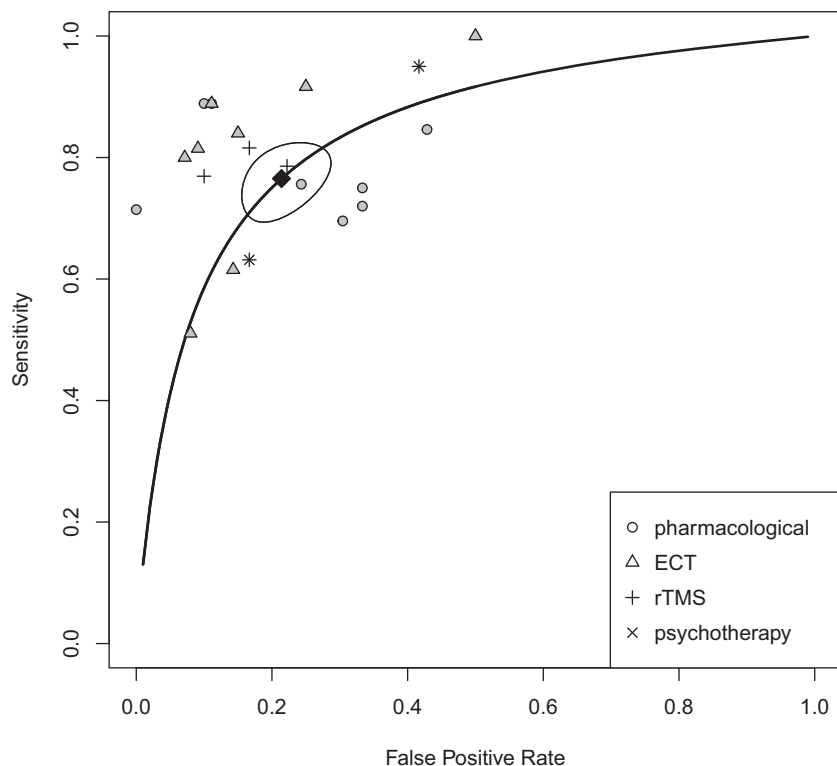


Fig. 2 Overall accuracy measures: area under the curve 0.84 (95% CI 0.81–0.87), sensitivity 77% (95% CI 71–82), specificity 79% (95% CI 73–84). Reitsma bivariate SROC model of the receiver operating characteristic curve. Summary of sensitivity and false-positive rate ($1 - \text{specificity}$) is indicated in black, sensitivity and false-positive rates for different interventions are gray-scale. ECT electroconvulsive therapy, rTMS repetitive transcranial magnetic stimulation, pharmacological pharmacotherapeutic antidepressive interventions.

used transcranial magnetic stimulation, and 8% used cognitive therapy. Most studies used either sMRI (31%) or task-based fMRI (31%), most often using emotional stimuli, 19% used resting-state fMRI, and 8% used DTI. Two studies combined multiple modalities^{40,50}.

As machine learning paradigm, 31% studies used support vector machine (SVM) for data-analysis, while 28% used logistic regression. After comparing classification accuracy with multiple algorithms (among others, SVM and random forest), Patel and colleagues used an alternating decision tree method⁵⁰. For validation, 85% used leave-one-out cross-validation. Two studies used an independent cohort to validate their results, while one study first cross-validated classification results, after which authors validated their prediction model in two small, independent cohorts, achieving similar results^{39,43,53}. For additional information on approaches to imaging analysis, please refer to Supplementary Table 2.

Meta-analysis

General outcome

After pooling results from studies with overlapping patient samples, we quantitatively analyzed 22 samples,

including one independent cohort replication that we have interpreted as a separate study⁴³. For all imaging modalities and interventions taken together, the meta-analytic estimate for the SROC AUC was 0.84 (95% CI 0.81–0.87), with 77% sensitivity (95% CI 71–82) and 79% specificity (95% CI 73–84), amounting to a moderately high classification performance (see Fig. 2).

Intervention differences

Sensitivity and specificity of ECT interventions were 80% (95% CI 73–85) and 83% (95% CI 72–90), respectively, compared to 75% (95% CI 68–82) and 72% (95% CI 64–80) for antidepressant medication. Exclusion of the studies that did not use SSRI as pharmacological agent had little influence on the results⁴⁹. Although prediction outcomes in ECT studies do show a trend toward higher precision, CIs overlapped (see Table 2). With only few primary studies, sensitivity and specificity for psychotherapy were, respectively, 84% (95% CI 68–92) and 72% (39–92), for repetitive transcranial magnetic stimulation (rTMS), respectively, 79% (95% CI 71–86) and 82% (74–88).

Modality differences

In order to assess whether sMRI studies yielded different performance measures compared to fMRI studies, we performed random-effect meta-regression for modality subtypes. When comparing fMRI and sMRI, z -regression values for sensitivities and specificities were non-significant, suggesting that prediction success for structural or functional neuroimaging did not differ between studies (see Table 3). Post hoc analysis excluding DTI and subdividing task-based and resting-state fMRI did not alter the results.

Quality assessment

Three studies included only late-life depression, which reduces applicability in the general MDD population (see Supplementary Fig. 1 and Supplementary Table 3). In terms of flow and timing, drop-outs were a common issue, with 10 studies having a drop-out rate of $\geq 30\%$, while 11 studies did not clarify drop-outs, possibly leading to attrition bias. Furthermore, two studies adapted the definition of response to create an even split in responders/non-responders, causing applicability concerns^{45,48}. One study did not pre-specify the pharmacological intervention⁵⁰.

Heterogeneity and publication bias

The univariate forest plot of diagnostic performance (in \ln OR) showed considerable overlap in CIs between studies with different ORs, indicating that heterogeneity might be caused by sample variance (see Fig. 3)²³. As described in the study description above, inter-study differences were present in population, modalities,

intervention type, response/remission definition, feature selection, and analysis technique. Deeks' funnel plot asymmetry test showed study size and diagnostic OR to be inversely related ($p = 0.044$; see Supplementary Fig. 2), indicating that classification performance was lower in studies with larger samples. Inspection of the gray literature that was excluded due to missing information in order to construct a confusion matrix (all of which were conference/poster abstracts) showed that the gray literature had comparable mean sample sizes ($n = 22$, mean $n = 56$) and accuracies (ranging from 73 to 95%) compared to the included studies. For an overview of gray literature results, see Supplementary Table 4.

Discussion

Our results show that machine learning analysis of MRI data can predict antidepressive treatment success with an AUC of 0.84, 77% sensitivity, and 79% specificity (Fig. 2). Furthermore, we did not find a difference in classification performance between studies using pharmacotherapy and ECT. Although ECT showed somewhat higher sensitivity and specificity, CIs largely overlapped between the two intervention types (Table 2). There were few primary studies for psychotherapy and rTMS, which also show overlapping CIs. In addition, classification performance of sMRI and fMRI did not differ significantly (Table 3).

To our knowledge, this is the first meta-analysis specifically examining MRI for predicting treatment effects in depression. The overall classification performance is comparable to the one reported by Lee et al., who found a general accuracy of 85% when combining the results for different neuroimaging modalities (defined as EEG, computed tomography, positron emission tomography, or MRI)⁵⁸. Those results were, however, based on a total of 8 MRI studies, whereas our search resulted in 22 individual studies for analysis. This is partly due to the time gap between studies, which underscores the rapid development in this research area. Our results show that MRI prediction studies perform somewhat better than EEG (AUC of 0.76) and comparable to accuracy of *diagnostic* classification studies with MRI that distinguishes depressed patients and healthy controls^{20,59}. In contrast to the review of EEG studies, we excluded studies that tested their model on the training set, which increased

Table 2 Summary estimates of sensitivity/specificity for different interventions.

Intervention group	Sensitivity	95% CI	Specificity	95% CI
Combined	77%	71–82	79%	73–84
Medication	75%	68–82	73%	64–80
ECT	80%	73–85	83%	72–90
Psychotherapy	84%	68–92	72%	39–92
rTMS	79%	71–86	82%	74–88

CI confidence interval, rTMS repetitive transcranial magnetic stimulation, ECT electroconvulsive therapy.

Table 3 Bivariate random-effect meta-regression z-scores for modality as covariate.

	Point estimate	Standard error	95% Lower	95% Upper	z-value	p Value
Sensitivity	0.221	0.233	-0.236	0.677	0.948	0.343
Specificity	0.217	0.252	-0.77	0.711	0.861	0.389

p Values for both sensitivity and specificity > 0.05 , i.e., z-score differences for functional and structural MRI are non-significant.

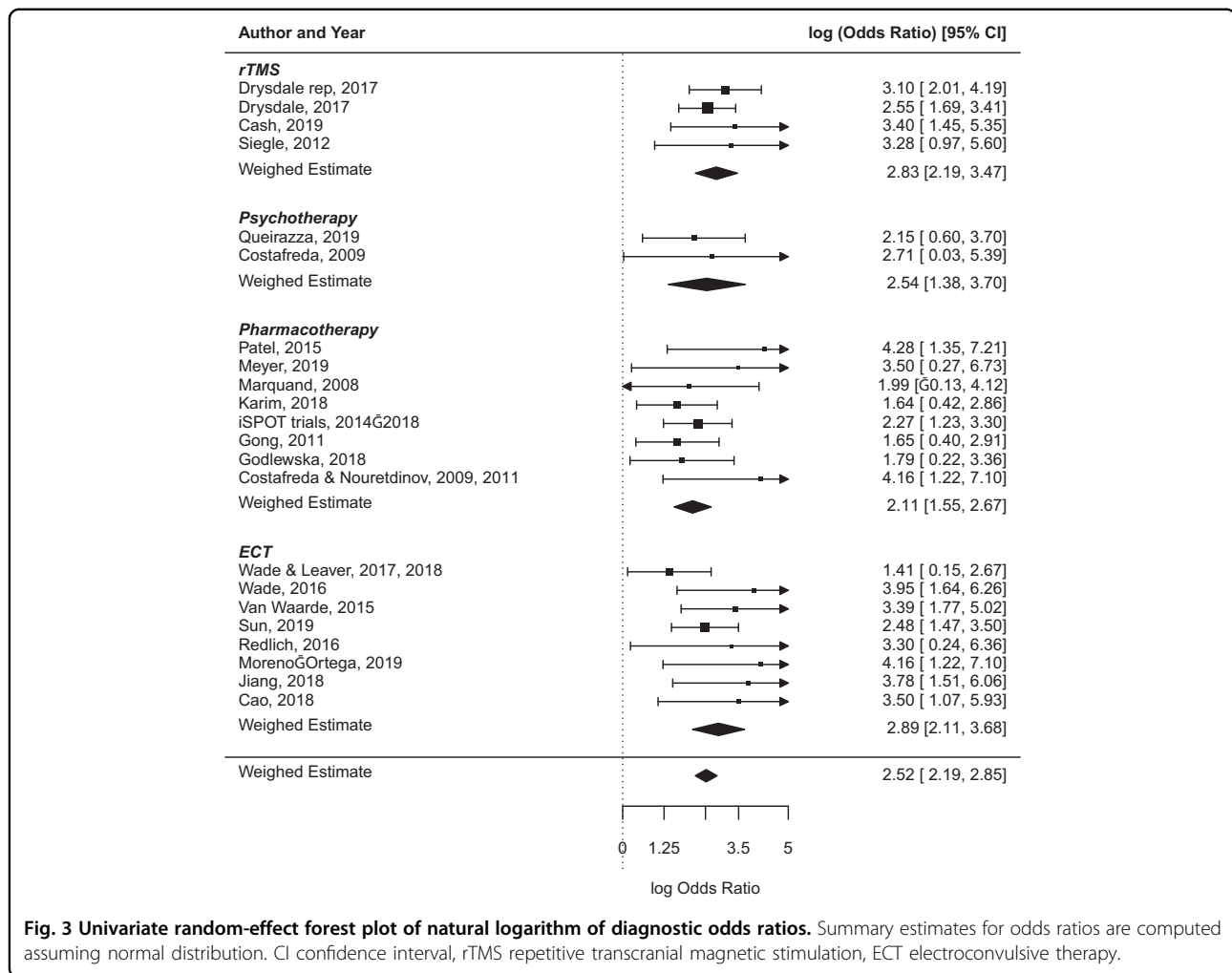


Fig. 3 Univariate random-effect forest plot of natural logarithm of diagnostic odds ratios. Summary estimates for odds ratios are computed assuming normal distribution. CI confidence interval, rTMS repetitive transcranial magnetic stimulation, ECT electroconvulsive therapy.

generalizability of our sample and avoided presenting inflated accuracy results.

Clinical practice would require different prediction approaches for a broad range of specific settings. It would be useful to have a single predictive test for therapy-resistant patients, especially to guide decision-making for invasive treatments such as ECT. For example, ECT is associated with cognitive side effects that are preferably avoided in case the treatment is unsuccessful⁶⁰. In addition, ECT is only applied in 1–2% of patients with persistent or severe depression and a biomarker that indicates a high probability of success may reduce the hesitance of its use⁶¹. However, for most treatments, a differential biomarker would be preferable, which would enable selecting the treatment with the highest chance of success. As of yet, no MRI study has used such prospective prediction and subsequent treatment matching to guide decision-making between two treatment options (for instance, between cognitive behavioral therapy and an SSRI). Furthermore, no studies have yet compared efficacy

of prediction-guided treatments versus regular treatment based on patient–clinician preference. Thus, although the predictive performance of MRI biomarkers is certainly promising, the current study designs do not yet enable the translation of research findings to the clinic.

Generally, studies were of acceptable quality, although drop-out rates could cause concern in terms of reliability. Drop-out rates were not mentioned in 11 studies, and for 10 studies, drop-out rates were >30% without using an intention-to-diagnose approach. Not accounting for drop-outs, who might be less likely to respond to treatment, could inflate response/remission data and consequently alter sensitivity and specificity of the predictive test. Additionally, our results show between-study variety regarding the response criterion, which typically consisted of clinical response ($\geq 50\%$ symptom reduction) or symptom remission. Different clinical settings might require different prediction outcomes. For instance, one could expect treatment of a first-time depressive episode to lead to complete remission, while in severe treatment-resistant

depression, response might be a more practical and achievable goal⁶². Authors should take care to pre-specify which outcome they will use and why that outcome is the most appropriate for their sample or intervention.

Furthermore, although no objective investigation for clinical heterogeneity in prediction studies exists, our random-effect forest plot shows considerable overlap of CIs with differing study results, implying the presence of sampling variation (Fig. 3)²². Clinical variance between samples is an important obstacle in generalizability of any diagnostic or predictive marker, especially in psychiatric illnesses such as MDDs, which is heterogeneous in both its clinical and neurophysiological manifestation^{63,64}. Thus, inter-sample diversity of inclusion criteria and methodological design might hamper the realization of a reliable predictive biomarker.

In the current literature on diagnostic accuracy studies, the possibility of publication selection as a source of bias is still under debate^{25,65}. Common forms of formalizations of publication bias, such as the Egger's or Begg's test, are not recommended for meta-analyses of prediction studies, since their sensitivity in diagnostic accuracy studies is generally poor²³. However, the recommended Deeks' funnel plot asymmetry test (see Supplementary Fig. 2) shows the presence of a sample size effect, with the n of a study being negatively correlated to classification performance, which could be attributable to publication bias⁶⁶. Another explanation of this significant correlation might be that large-scale studies with large samples are more likely to consist of heterogeneous patient groups, which in turn reduces prediction accuracy⁶⁷. As a further exploration of publication bias, our search also took into account gray literature, which indicated that publication (or positive result) bias was absent. In conclusion, quantitative testing could not distinguish between a real effect (due to accuracy reduction in large heterogeneous samples) or publication bias. Although the gray literature deems its presence less likely, we cannot exclude the presence of publication bias.

The following limitations warrant further discussion. First, we did not find modality differences, but studies conducting fMRI research might have also attempted prediction with (less time-consuming and cheaper) sMRI, which remained unpublished. Although we did contact authors for additional information, response was poor, so we were unable to rule out reporting bias for modality differences. We would advise authors of future studies to publish non-significant results as well as significant but less accurate results, since both are potentially useful in comparing the merits of different modalities. Second, the number of studies predicting the effects of psychotherapy, specifically cognitive therapy, outcome was low, resulting in a blind spot for a commonly deployed treatment in MDD⁶⁸. Third, cross-validation in small samples results in

large variation of the estimated accuracy, and as indicated above, accuracy reduces with larger sample heterogeneity^{67,69}. Since the mean sample size of our studies was 44 (with a median n of 33), the reported results may be optimistic because of overfitting. Overfitting is a cause for concern specifically in MRI studies, with relatively small sample sizes and large amounts of fitted data⁷⁰. Furthermore, characteristics of the test set during cross-validation will approximate the characteristics of the training set more than when tested in the general population, due to selection bias⁷¹. Only two included studies replicated their training data in an independent cohort, and one included study used an out-of-sample cohort to further test their cross-validated results, leaving the question open to which extent the majority of results can be generalized to new patients.

In order to optimize patient care, reduce treatment resistance, and shorten duration of illness, developing models that predict treatment success on individual-patient level is an urgent task. In a 2012 consensus report on diagnostic imaging markers in psychiatry, the American Psychiatric Association research council proposed 80% sensitivity and specificity as prerequisite for the clinical application of a biomarker⁷². Furthermore, biomarkers should be ideally be reliable, reproducible, non-invasive, simple to perform, and inexpensive. The results for an ECT biomarker fulfilled the 80% criterion, but the results for a medication biomarker fell short. But following these terms, primarily reproducibility has not yet been sufficiently well established with small sample sizes and external validation in only a minority of studies. This precludes recommending MRI for treatment response prediction in clinical practice at this point. Future multicenter studies with large patient samples that represent clinical heterogeneity are required to warrant MRI biomarker generalizability⁷³. However, one might question whether excellent generalizability is a goal that should be aimed for: if each clinical site were to develop its own locally reliable and replicable biomarker that incorporates the local hardware, patient, and treatment variability, the predictive accuracy is expected to be higher than when all potential sources of heterogeneity are accounted for^{67,74}. Standard machine learning analysis would, then, mean a departure from the traditional universalist paradigm in diagnostics and instead initiate a shift to a paradigm of localization: heterogeneous yet locally applicable classification models. This will enable to retrain predictive models to obtain even better performance with more data after biomarker deployment. And this may enable to take advantage rather than disadvantage from (inevitable) hardware upgrades, such as higher signal-to-noise for new generations of MR scanners and coils.

In conclusion, prediction of treatment success using machine learning analysis of MRI data holds promise but

has not transcended the research status and should not yet be implemented into clinical practice. Once it overcomes the aforementioned hurdles, MRI may become a clinical decision support tool aimed to reduce unsuccessful treatments and improve treatment efficacy and efficiency.

Acknowledgements

We thank Joost Daams, clinical librarian at the Amsterdam UMC hospital, for his excellent help in defining our search terms and in running the search. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author details

¹Department of Psychiatry, Amsterdam Neuroscience, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. ²Department of Child and Adolescent Psychiatry, Amsterdam University Medical Center, Amsterdam, The Netherlands. ³Centre for Urban Mental Health, Amsterdam University Medical Center, Amsterdam, The Netherlands

Conflict of interest

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41398-021-01286-x>.

Received: 29 October 2020 Revised: 5 February 2021 Accepted: 18 February 2021

Published online: 15 March 2021

References

- Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).
- Kawakami, N. et al. Early-life mental disorders and adult household income in the World Mental Health Surveys. *Biol. Psychiatry* **72**, 228–237 (2012).
- Kessler, R. C. & Bromet, E. J. The epidemiology of depression across cultures. *Annu. Rev. Public Health* **34**, 119–138 (2013).
- Hasin, D. S. et al. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry* **75**, 336–346 (2018).
- Herrman, H. et al. Reducing the global burden of depression: a Lancet–World Psychiatric Association Commission. *Lancet* **393**, e42–e43 (2019).
- Gelenberg, A. J. et al. American Psychiatric Association Practice Guideline for the Treatment of Patients With Major Depressive Disorder, Third Edition. *Am. J. Psychiatry* **167**, 167 (2010).
- Pigott, H. E., Leventhal, A. M., Alter, G. S. & Boren, J. J. Efficacy and effectiveness of antidepressants: current status of research. *Psychother. Psychosom.* **79**, 267–279 (2010).
- Loerinc, A. G. et al. Response rates for CBT for anxiety disorders: need for standardized criteria. *Clin. Psychol. Rev.* **42**, 72–82 (2015).
- John Rush, A. et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am. J. Psychiatry* **163**, 1905–1917 (2006).
- Heijnen, W. T., Birkenhager, T. K., Wiersma, A. I. & van den Broek, W. W. Antidepressant pharmacotherapy failure and response to subsequent electroconvulsive therapy: a meta-analysis. *J. Clin. Psychopharmacol.* **30**, 616–619 (2010).
- Kellner, C. H. et al. ECT in treatment-resistant depression. *Am. J. Psychiatry* **169**, 1238–1244 (2012).
- McIntyre, R. S. et al. Treatment-resistant depression: definitions, review of the evidence, and algorithmic approach. *J. Affect. Disord.* **156**, 1–7 (2014).
- Kennis, M. et al. Prospective biomarkers of major depressive disorder: a systematic review and meta-analysis. *Mol. Psychiatry* **25**, 321–338 (2020).
- Ozomaro, U., Wahlestedt, C. & Nemeroff, C. B. Personalized medicine in psychiatry: problems and promises. *BMC Med.* **11**, 132 (2013).
- Perlman, K. et al. A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *J. Affect. Disord.* **243**, 503–515 (2019).
- Bzdok, D. & Meyer-Lindenberg, A. Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3**, 223–230 (2018).
- Yahata, N., Kasai, K. & Kawato, M. Computational neuroscience approach to biomarkers and treatments for mental disorders. *Psychiatry Clin. Neurosci.* **71**, 215–237 (2017).
- Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930 (2015).
- Lee Y. et al. Corrigendum to “Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review”. *J. Affect. Disord.* **241** (2018) 519–532. *J. Affect. Disord.* **274**, 1211–1215 (2020).
- Widge, A. S. et al. Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. *Am. J. Psychiatry* **176**, 44–56 (2019).
- Reitsma, J. B. et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **58**, 982–990 (2005).
- Macaskill, P., Gatsonis, C., Deeks, J. J., Harbord, R. M. & Takwoingi, Y. in *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0* (eds Deeks, J. J., Bossuyt, P. M. & Gatsonis, C.) Ch. 10 (The Cochrane Collaboration, 2010).
- Bossuyt, P. D. C., Deeks, J., Hyde, C., Leeflang, M. & Scholten, R. (eds) *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0* (The Cochrane Collaboration, 2010).
- Deeks, J. J., Macaskill, P. & Irwig, L. 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 (2005).
- van Enst, W. A., Ochodo, E., Scholten, R. J. P. M., Hooft, L. & Leeflang, M. M. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. *BMC Med. Res. Methodol.* **14**, 70–70 (2014).
- Whiting, P. F. et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**, 529–536 (2011).
- Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).
- Doebler, P. mada: meta-analysis of diagnostic accuracy. R package version 0.5.10. <https://rdrr.io/forge/mada/> (2020).
- R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
- Costafreda, S. G., Chu, C., Ashburner, J. & Fu, C. H. Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS ONE* **4**, e6353 (2009).
- Costafreda, S. G., Khanna, A., Mourao-Miranda, J. & Fu, C. H. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. *Neuroreport* **20**, 637–641 (2009).
- Nouretdinov, I. et al. Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. *Neuroimage* **56**, 809–813 (2011).
- Siegle, G. J. et al. 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 (2012).
- Queirazza, F., Fouragnan, E., Steele, J. D., Cavanagh, J. & Philiastides, M. G. Neural correlates of weighted reward prediction error during reinforcement learning classify response to cognitive behavioral therapy in depression. *Sci. Adv.* **5**, eaav4962 (2019).
- van Waarde, J. A. et al. A functional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. *Mol. Psychiatry* **20**, 609–614 (2015).
- Moreno-Ortega, M. et al. Resting state functional connectivity predictors of treatment response to electroconvulsive therapy in depression. *Sci. Rep.* **9**, 5071 (2019).
- Sun, H. et al. Preliminary prediction of individual response to electroconvulsive therapy using whole-brain functional magnetic resonance imaging data. *Neuroimage Clin.* **26**, 102080 (2020).

38. Redlich, R. et al. Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. *JAMA Psychiatry* **73**, 557–564 (2016).
39. Jiang, R. et al. SMRI biomarkers predict electroconvulsive treatment outcomes: accuracy with independent data sets. *Neuropsychopharmacology* **43**, 1078 (2017).
40. Leaver, A. M. et al. Fronto-temporal connectivity predicts ECT outcome in major depression. *Front. Psychiatry* **9**, 92 (2018).
41. Wade, B. S. C. et al. Data-driven cluster selection for subcortical shape and cortical thickness predicts recovery from depressive symptoms. *Proc. IEEE Int. Symp. Biomed. Imaging* **2017**, 502–506 (2017).
42. Cao, B. et al. Predicting individual responses to the electroconvulsive therapy with hippocampal subfield volumes in major depression disorder. *Sci. Rep.* **8**, 5434 (2018).
43. Drysdale, A. T. et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* **23**, 28–38 (2017).
44. Cash, R. F. H. et al. A multivariate neuroimaging biomarker of individual outcome to transcranial magnetic stimulation in depression. *Hum. Brain Mapp.* **40**, 4618–4629 (2019).
45. Gong, Q. et al. Prognostic prediction of therapeutic response in depression using high-field MR imaging. *Neuroimage* **55**, 1497–1503 (2011).
46. Marquand, A. F., Mourao-Miranda, J., Brammer, M. J., Cleare, A. J. & Fu, C. H. Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. *Neuroreport* **19**, 1507–1511 (2008).
47. Godlewska, B. R. et al. Predicting treatment response in depression: the role of anterior cingulate cortex. *Int. J. Neuropsychopharmacol.* **21**, 988–996 (2018).
48. Meyer, B. M. et al. Prefrontal networks dynamically related to recovery from major depressive disorder: a longitudinal pharmacological fMRI study. *Transl. Psychiatry* **9**, 64 (2019).
49. Karim, H. T. et al. Acute trajectories of neural activation predict remission to pharmacotherapy in late-life depression. *Neuroimage Clin.* **19**, 831–839 (2018).
50. Patel, M. J. et al. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *Int. J. Geriatr. Psychiatry* **30**, 1056–1067 (2015).
51. Goldstein-Piekarski, A. N. et al. Human amygdala engagement moderated by early life stress exposure is a biobehavioral target for predicting recovery on antidepressants. *Proc. Natl Acad. Sci. USA* **113**, 11955–11960 (2016).
52. Goldstein-Piekarski et al. Intrinsic functional connectivity predicts remission on antidepressants: a randomized controlled trial to identify clinically applicable imaging biomarkers. *Transl. Psychiatry* **8**, 57 (2018).
53. Grieve, S. M., Korgaonkar, M. S., Gordon, E., Williams, L. M. & Rush, A. J. Prediction of nonremission to antidepressant therapy using diffusion tensor imaging. *J. Clin. Psychiatry* **77**, e436–e443 (2016).
54. Korgaonkar, M. S., Williams, L. M., Song, Y. J., Usherwood, T. & Grieve, S. M. Diffusion tensor imaging predictors of treatment outcomes in major depressive disorder. *Br. J. Psychiatry* **205**, 321–328 (2014).
55. Williams, L. M. et al. Amygdala reactivity to emotional faces in the prediction of general and medication-specific responses to antidepressant treatment in the randomized iSPOT-D trial. *Neuropsychopharmacology* **40**, 2398–2408 (2015).
56. Wade, B. S. et al. Effect of electroconvulsive therapy on striatal morphometry in major depressive disorder. *Neuropsychopharmacology* **41**, 2481–2491 (2016).
57. Williams, L. M. et al. International study to predict optimized treatment for depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials* **12**, 4–4 (2011).
58. Lee, Y. et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *J. Affect. Disord.* **241**, 519–532 (2018).
59. Kambeitz, J. et al. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. *Biol. Psychiatry* **82**, 330–338 (2017).
60. Semkowska, M. & McLoughlin, D. M. Objective cognitive performance associated with electroconvulsive therapy for depression: a systematic review and meta-analysis. *Biol. Psychiatry* **68**, 568–577 (2010).
61. Slade, E. P., Jahn, D. R., Regenold, W. T. & Case, B. G. Association of electroconvulsive therapy with psychiatric readmissions in US hospitals. *JAMA Psychiatry* **74**, 798–804 (2017).
62. Rush, A. J. et al. Report by the ACNP Task Force on Response and Remission in Major Depressive Disorder. *Neuropsychopharmacology* **31**, 1841–1853 (2006).
63. Fried, E. I. Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert Rev. Neurother.* **17**, 423–425 (2017).
64. Dinga, R. et al. Evaluating the evidence for biotypes of depression: methodological replication and extension of Drysdale et al. *Neuroimage Clin.* **2019**, 101796 (2017).
65. Murad, M. H., Chu, H., Lin, L. & Wang, Z. The effect of publication bias magnitude and direction on the certainty in evidence. *BMJ Evid. Based Med.* **23**, 84 (2018).
66. Leeflang, M. M. G. Systematic reviews and meta-analyses of diagnostic test accuracy. *Clin. Microbiol. Infect.* **20**, 105–113 (2014).
67. Schnack, H. G. & Kahn, R. S. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Front. Psychiatry* **7**, 50–50 (2016).
68. Widnall, E., Price, A., Trompeter, H. & Dunn, B. D. Routine cognitive behavioural therapy for anxiety and depression is more effective at repairing symptoms of psychopathology than enhancing wellbeing. *Cogn. Ther. Res.* **44**, 28–39 (2020).
69. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77 (2018).
70. Mateos-Pérez, J. M. et al. Structural neuroimaging as clinical predictor: a review of machine learning applications. *Neuroimage Clin.* **20**, 506–522 (2018).
71. Schmidt, R. L. & Factor, R. E. Understanding sources of bias in diagnostic accuracy studies. *Arch. Pathol. Lab. Med.* **137**, 558–565 (2013).
72. First M. B. et al. *Consensus Report of the APA Work Group on Neuroimaging Markers of Psychiatric Disorders*. APA Council on Research Consensus Paper (APA, 2012).
73. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365 (2017).
74. Dluhos, P. et al. Multi-center machine learning in imaging psychiatry: a meta-model approach. *Neuroimage* **155**, 10–24 (2017).