SHORT TITLE: Test-retest reliability in fMRI of depression

Importance of test-retest reliability for promoting fMRI based screening and interventions in major depressive disorder

Laurie Compère, PhD [1], Greg J. Siegle, PhD [1], Kymberly Young, PhD [1]

[1] Department of Psychiatry, University of Pittsburgh School of Medicine

Western Psychiatric Institute and Clinic

3811 O'Hara Street

Pittsburgh, PA 15213

ljc44@pitt.edu

gsiegle@pitt.edu

youngk@pitt.edu

Corresponding author

Laurie Compère, Department of Psychiatry, University of Pittsburgh School of Medicine

Western Psychiatric Institute and Clinic

3811 O'Hara Street

Pittsburgh, PA 15213

Email: ljc44@pitt.edu

Abstract

Proponents of personalized medicine have promoted neuroimaging evaluation and treatment of major depressive disorder in three areas of clinical application: clinical prediction, outcome evaluation, and neurofeedback. Whereas psychometric considerations such as test-retest reliability are basic precursors to clinical adoption for most clinical instruments, they are often not considered for neuroimaging assessments. As an example, we consider functional magnetic resonance imaging (fMRI) of depression, a common and particularly well validated mechanistic technology for understanding disorder and guiding treatment. In this article, we review work on test-retest reliability for depression fMRI studies. We find that basic psychometrics have not been regularly attended to in this domain. For instance, no fMRI neurofeedback study has included measures of test-retest reliability despite the implicit

assumption that brain signals are stable enough to train. We consider several factors that could be useful to aid clinical translation including 1) attending to how the BOLD response is parameterized, 2) identifying and promoting regions or voxels with stronger psychometric properties 3) accounting for within-individual changes (e.g., in symptomatology) across time and 4) focusing on tasks and clinical populations that are relevant for the intended clinical application. We apply these principles to published prognostic and neurofeedback data sets. The broad implication of this work is that attention to psychometrics is important for clinical adoption of mechanistic assessment, is feasible, and may improve the underlying science.

Keywords: depression, fMRI, neurofeedback, psychometric, treatment prediction, test-retest reliability

1. GENERAL INTRODUCTION

Proponents of personalized medicine have promoted mechanistic evaluation and mechanistically targeted treatments for major depressive disorder (Hansen and Siegle, 2015). As an example, we consider functional magnetic resonance imaging (fMRI), a common and particularly well validated mechanistic technology that represents a promising proof-of-concept in this area. Longitudinal assessment of changes in regional brain activity using functional magnetic resonance imaging (fMRI) has increasingly been used in research on the treatment of psychiatric conditions including major depressive disorder (MDD) (Fournier et al., 2014). As good psychometric properties are essential for any measure to be considered for clinical adoption (Pickford and Guilford, 2007), best-practice guidelines for increasing generalizability and reproducibility of fMRI results are emerging (Nichols et al., 2017; Poldrack et al., 2017). We focus here on test-retest reliability in task-based fMRI and neurofeedback (fMRI-nf) designs, using MDD as a running case example. Ideally, our observations can be applied to other technologies and across neuropsychiatric disorders.

To understand the current state of the field, we conducted literature reviews quantifying how often test-retest reliability was reported in fMRI biomarker and real-time fMRI neurofeedback (rtfMRI-nf) studies in MDD. As we will demonstrate below, this was infrequent and the general literature has shown that wen assessed, reliability was generally low. We thus suggest a few analytic techniques for improving test-rest reliability in fMRI and its clinical applicability. We focus on data analysis to make our suggestions maximally applicable to already collected data. Finally, we test these suggested principles on published MDD neuroimaging treatment outcome and neurofeedback datasets as proofs of concept.

The idea that fMRI could have therapeutic utility is based on assumptions that hemodynamic activity is reliable over time in the absence of intervention, and that observed changes between one scan session and the next have significant and interpretable values (Barch and Mathalon, 2011). The reliability of fMRI also affects its criterion validity, as poor reliability limits the strength of association between the instrument and other relevant measures (Vul et al., 2009).

1.1. On computing reliability of fMRI

Demonstrating ability to achieve similar results over time, or the reliability of measures is considered critical to creating a clinically useful measure (Pickford and

Guilford, 2007). Reliability is a quantitative measure of stability of an individual's data (Bennett and Miller, 2013). It refers to the ability of a measure to distinguish participants from each other and to replicate the order of individuals' ranks during repeated assessments, assuming they do not experience true signal change between assessments (Barch and Mathalon, 2011).

Though stable regional hemodynamic activations at the group level can be observed over time, there are significant changes in how each subject contributes individually to the observed group activation (Caceres et al., 2009; Zandbelt et al., 2008). Various approaches have been used to measure test-retest reliability for fMRI. For example, a Pearson correlation between visits across time measure the degree to which visits on two occasions are linearly related, where data from each visit are independently scaled (e.g., Harrington et al., 2006). A more common approach, and the measure we focus on in this manuscript, involves computing intra-class correlation coefficients (ICC) that also reflect rank ordering of values across days (Bennett and Miller, 2010) as a ratio of variance between values observed across subjects and sites divided by the total visit variance (Bartko, 1966). Values range from 0 (no reliability) to 1 (perfect reliability). There are three different types of ICCs described by the princeps article written by Shrout and Fleiss (1979). The ICC(1,1) index is similar to the Pearson correlation but normalizes by the pooled mean and variance across visits. ICC(2,1) is an agreement index that allows generalization of results across scanners while ICC(3,1) works under the assumption that the variance is the same across scanners. Therefore, the ICC(3,1), mostly used across studies, is a scanner consistency index where the effect of scanner is considered a fixed effect (Shrout and Fleiss, 1979). In order to match the literature in the field and because we considered the scanner as a covariate of interest when investigating the impact of taking into account clinical and design covariates when computing reliability indexes, we mainly used ICC(3,1) in our analyses.

Interpretation of ICC values is subjective with no uniformly accepted standards; ICC values of less than 0.4 are often considered poor, 0.4-0.59 fair, 0.60-0.74 good, and above 0.75 excellent (Cicchetti, 1994; Plichta et al., 2012; Shrout and Fleiss, 1979), though more stringent cutoffs have also been recommended (e.g., Portney and Watkins, 2009). Negative ICC value are usually interpreted as no reliability (Bartko, 1976), since these values are outside the theoretical limits of ICC (although negative

values may appear when within-subject variance is greater than between-subject variance) (Lahey et al., 1983).

Though the ICC has been recommended for use in fMRI (Caceres et al., 2009), some fMRI analysis packages (SPM, FSL) do not inherently support computation of this metric, potentially hinting at its perceived value in the field, though other packages (e.g., AFNI, NIfTI) do provide for its computation, and add-on packages (e.g., reliability toolbox for SPM or other packages on R) do allow such computations (see Computation of voxelwise ICCs using different tools in Box 3 in supplementary materials for more details). Indeed, reliability estimates have been rarely reported in fMRI studies and usually reveal poor reliability when estimated (Elliott et al., 2020). Non-clinical studies have generally found low to moderate test-retest reliability values for regional fMRI activity, with ICCs ranging from 0.33-0.66 (reviewed in Bennett and Miller, 2010).

1.2. Biomarker/Prediction Studies Review

Many studies suggest fMRI measurements can be used to predict treatment outcome in MDD (for reviews, see Arnone, 2019; Fonseka et al., 2018; Phillips and Swartz, 2014; Wessa and Lois, 2015). The underlying assumption is that biomarkers in the brain are involved in the causal process of MDD. Therefore, it is expected that the activity measured in these biomarkers is related to, and evolves over time with, symptom changes in general and that for interventions targeting the biomarker the more abnormal activity observed, the more effective the intervention will be. However, clinical applications of these findings are limited by the possibility that these biomarkers may have low test-retest reliability (Nord et al., 2017). If a biomarker is not reliable, it is impractical to interpret its activation at the individual level (Fu et al., 2013; Guo et al., 2012). Thus, despite strong predictive utility, researchers acknowledge that their results might be limited by poor test-retest reliability (e.g., Fu et al., 2015). Of particular interest, the amygdala, a commonly reported biomarker for MDD, shows poor to good reliability when emotional stimuli are displayed, with great heterogeneity between studies in healthy participants (Lois et al., 2018). Thus, we surveyed the predictive fMRI literature in MDD to examine whether this first step was being taken.

1.2.1. Method

A PubMed search with the key words “fMRI AND biomarker OR prediction OR predict AND depression OR MDD OR major depressive disorder NOT Rest NOT

Resting" produced 140,640 results. We combined this list with other articles discovered in our submitted fMRI meta-analysis of depression treatment outcome prediction studies (Strege et al., 2020) to complete the list of articles (Table 1)." After removing articles not including functional neuroimaging (i.e., studies focusing on volumetric measures or using PET) or human participants, we were left with 55 studies (Table 1).

Table 1: Studies Examining neuroimaging biomarkers of pharmacotherapy and psychotherapy outcomes in Major Depressive Disorder and mention of test-retest reliability of the studies

Reference	Treatment(s)	Biomarker	Findings	Mention of signal reliability	Possibility to test signal reliability
Sheline et al.(2001)	Sertraline	Amygdala	Decreased activation following treatment	No	Yes
Davidson, et al. (2003)	Venlafaxine	ACC	Greater activation at baseline associated with better treatment response	No	Yes
Fu et al. (2004)	Fluoxetine	ACC, ventral striatum, cerebellum	Reduction of dynamic range associated with symptoms improvement	No	Yes
Canli et al. (2005)	None	Amygdala	Amygdala activation at baseline predicts symptom improvement	No	No

Schaefer, et al. (2006)	Venlafaxine	Prefrontal, temporal and parietal cortices, insula, basal ganglia and hippocampus	Normalized activation after treatment	No	Yes
Siegle, Carter, & Thase (2006)	CBT	sgACC and amygdala	Low and high, respectively, activation is associated with greater symptom improvement after therapy	No	No
Anand, et al. (2007)	Sertaline	Amygdala and ACC	Decrease activation in limbic regions and increased connectivity with the ACC after treatment	No	Yes
Chen et al. (2007)	Fluoxetine	ACC	Greater activation at baseline predict faster rates of symptom improvement	No	No
Fales et al. (2007)	Escitalopram	DLPFC	Enhanced activation following treatment	No	Yes
Fitzgerald et al. (2007)	TMS	Middle frontal gyrus, left precuneus, left precentral gyrus, left medial frontal	Decreased activation after low frequency treatment in middle frontal gyrus and left	No	Yes

		gyrus, right inferior frontal gyrus	precuneus in respondents – Increased activation after high frequency treatment in left prefrontal gyrus, left medial frontal gyrus, right inferior frontal gyrus in respondents		
Fu et al. (2007)	Fluoxetine	Hippocampus and extrastriate cortex	Greater activation following treatment and associated with symptom improvement	No	Yes
Langenecker et al. (2007)	S-citalopram	Insula, right middle frontal gyrus, left inferior frontal gyrus, amygdala and cerebellar vermis	Greater activation at baseline associated with symptoms improvement	No	Yes
Robertson et al. (2007)	Bupropion	Amygdala	Reduced activation associated with symptom improvement	No	Yes
Walsh et al. (2007)	Fluoxetine	dACC, left middle frontal and lateral temporal cortices	Recuced activity at baseline associated with	Yes (discussion section)[a]	Yes

			symptom improvement		
Fu et al. (2008)	CBT	dACC	Reduced activation at baseline associated with symptom improvement	No	Yes
Benedetti et al. (2009)	Venlafaxine	Right medial frontal gyrus	Decreased activation following treatment was associated with symptom improvement	No	Yes
Costafreda, et al. (2009)	CBT	ACC, superior and middle frontal cortices, paracentral cortex, superior parietal cortex, precuneus and cerebellum	Activation contributed to prediction of remission	No	No
Dichter et al. (2010)	Behavioral Action Therapy	Paracingulate gyrus	Activation was prognostic for depressive symptom change after psychotherapy	No	Yes
Forbes et al., 2010	CBT and SSRI	Striatum and mPFC	Final levels of severity symptoms were related to pretreatment striatal reactivity	No	No

			and greater striatal and lower mPFC activity was prognostic for anxiety symptom reduction		
Keedwell et al. (2010)	Various antidepressants	Right visual cortex and right sgACC	Greater baseline activity associated with clinical improvement after treatment	No	Yes
Lemogne et al. (2010)	Various antidepressants	Left DLPFC	Reduced activation following treatment	No	Yes
López-Solà et al. (2010)	Duolexetine	pACC, right prefrontal cortex, pons	Clinical improvement associated with reduced activation	No	Yes
Roy et al. (2010)	Citalopram	Ventromedial prefrontal cortex and ACC	Greater activation at baseline associated with symptom improvement	No	Yes
Victor, et al. (2010)	Sertaline	Amygdala	Decreased activation after treatment	No	Yes
Wagner et al. (2010)	Citalopram, reboxetine	Amygdala, hippocampus	Decreased activation after citalopram treatment	No	Yes
Frodl et al. (2011)	Mirtazapine, venlafaxine	Left fusiform gyrus, right	Increased activation in the	No	Yes

		rolandic operculum	left fusiform gyrus at baseline was associated with a better response to venlafaxine and smaller activation in the right rolandic operculum was related to better response to mirtazapine		
Light et al. (2011)	Venlefaxine, fluoxetine	Ventrolateral prefrontal cortex	Reduced activity at baseline is associated with anhedonia reduction	No	Yes
Ritchey, et al. (2011)	CBT	Ventromedial prefrontal cortex	Increased activity at baseline associated with symptom improvement	No	Yes
Samson et al. (2011)	Mirtazapine, venlafaxine	dmPFC, posterior cingulate cortex, superior frontal gyrus, caudate nucleus and insula	Greater activation at treatment associated with better treatment response	No	Yes
Arnone et al. (2012)	Citalopram	Amygdala	Reduced activation	No	Yes

			following treatment		
Godlewska, et al. (2012)	Escitalopram	Amygdala	Reduced activity after treatment	No	No
Rosenblau et al. (2012)	Escitalopram	Amygdala, prefrontal cortex	Decreased activation following treatment	No	Yes
Ruhé, et al. (2012)	Paroxetine	Amygdala	Lower activation associated with better response to treatment after	No	Yes
Siegle et al. (2012)	CBT	sgACC	Reduced activation at baseline associated with greater symptom improvement	Yes	Yes
Stoy et al. (2012)	Escitalopram	Ventral striatum	Increased activation following treatment	No	Yes
Tao et al. (2012)	Fluoxetine	Amygdala, orbitofrontal cortex and sgACC	Decreased activation after treatment	Yes (discussion section)[b]	Yes
Wang et al. (2012)	Fluoxetine	Insula, left ACC and middle frontal gyrus	Decreased activation in insula and left ACC and greater in the middle frontal gyrus following treatment	No	Yes

Furey et al. (2013)	Scopolamine	Middle occipital cortex	Increased activation at baseline was prognostic for symptoms improvement	No	Yes
Heller et al. (2013)	Fluoxetineor venlafaxine	Nucleus accumbens	Greater activation following treatment associated with more self-reported positive affect	No	Yes
Miller et al. (2013)	Escitalopram	Midbrain, DLPFC, paracingulate, ACC, thalamus and caudate nuclei	Reduced activation at baseline correlated with greater improvement following treatment	No	No
Rizvi et al. (2013)	Fluoxetine and olanzapine	Premotor cortex	Increased activation at baseline in respondents was prognostic for symptom improvement	Yes but not reported (method section)[c]	Yes
Victor, et al. (2013)	Sertraline	pgACC	Increased correlation at baseline correlated with greater clinical improvement after treatment	No	Yes
Toki et al. (2014)	Various antidepressants	Left hippocampus	Increased activation	No	No

			associated with greater response treatment		
Yoshimura et al. (2014)	CBT	vACC	Improvements in depressive symptoms were negatively correlated with its activity	No	Yes
Fu et al. (2015)	Duloxetine	Posterior cingulate cortex	Increased activation following treatment	Yes (limitation section)[d]	Yes
Furey et al. (2015)	Scopolamine	sgACC and middle occipital cortex	Increased and decreased activation, respectively, associated with treatment response	No	Yes
Straub et al., 2015	CBT	sgACC	Activation before treatment related to therapeutic success	No	Yes
Williams et al. (2015)	Escitalopram, sertraline, venlafaxine	Amygdala	Decreased activation at baseline was associated with treatment response	No	Yes
Cullen et al. (2016)	Various antidepressants	Rostral and sgACC, insula, middle frontal cortex, right	Decreased activation in postral and sgACC and increased in ,	No	Yes

		hippocampus and left cerebellum	insula, middle frontal cortex, right hippocampus and left cerebellum associated with symptom improvement		
Delaveau et al. (2016)	Agomelatine	DLPFC and precuneus	Activation at baseline was related to treatment response	No	Yes
Doerig et al. (2016)	CBT	Amygdala	Activity in this region pre-intervention is negatively correlated with the outcome	No	No
Godlewska, et al. (2016)	Escitalopram	ACC, insula, amydgala and thalamus	Reduced activity after treatment associated with treatment response	No	Yes
Gyurak et al. (2016)	Escitalopram, sertraline and venlafaxine	DLPFC and inferior parietal cortex	Increased DLPFC activation at baseline associated with remission and increased inferior parietal activation associated with remission for SSRI and the	No	Yes

			opposite pattern for SNRI		
Opmeer et al. (2016)	-	Rostral ACC	Increased activation at baseline was prognostic for remission	No	Yes
Szczepanik et al. (2016)	Scopolamine	Amygdala	Increased activity at baseline was associated with symptoms improvement	Yes (limitation section)[e]	No
Fang et al., (2017)	Transcutaneous vagus nerve stimulation	Insula	Activation level at first stimulation session associated with clinical improvement	No	No
Sankar, et al. (2017)	Duloxetine	Left inferior frontal activity	Decreased activation following treatment	No	Yes
Spies et al. (2017)	Escitalopram	Precuneus and PCC	Deactivation before treatment was related to change in symptoms after 2 weeks of treatment	No	No
Godlewska et al. (2018)	Escitalopram	pgACC	Activity before treatment was able to predict response status (responder vs non-responder)	No	No

			at the level of individual participant		
Rubin-Falcone et al. (2018)	CBT	sgACC, medial prefrontal cortex, lingual gyrus	Increased activation following treatment associated with better treatment outcome	No	Yes

ACC: Anterior Cingulate Cortex; CBT: Cognitive Behavioral Therapy; dACC: dorsal Anterior Cingulate Cortex; DLPFC: DorsoLateral Prefrontal Cortex; dmPFC: dorsomedial Prefrontal Cortex; mPFC : medial Prefrontal Cortex ; MDD: Major Depressive Disorder; PCC : Posterior Cingulate Cortex; pgACC: pregenual Anterior Cingulate Cortex; sgACC: subgenual Anterior Cingulate Cortex ; SSRI : selective serotonine reuptake inhibitor

[a] "Test-retest effects were accounted for by the healthy control group, who underwent the same scans at the same time points"

[b] "repeat fMRI assessment of healthy comparison subjects, as well as repeat assessment of the depressed adolescents, thus providing assessment of expected test-retest reliability"

[c] "For analyses of change over time, a higher level fixed effects analysis was run for each subject, contrasting parameter estimates within subject for the response to slides at the two time points of interest."

[d]"perhaps in part reflecting the poor test-retest reliability of amygdala response to these emotional faces [54], while resting-state fMRI data show greater robustness and reproducibility [55]. Test-retest reliability of a neuroimaging measure becomes particularly important in the development of biomarkers for prognosis and diagnosis [44]."

[e]"some investigators have raised concerns regarding the reliability of the BOLD signal (Boubela et al., 2015). Nevertheless, studies have found that emotional stimuli evoke a consistent pattern of responsivity over repeated sessions (Johnstone et al., 2005)."

1.2.2. Results

Though most of the reviewed studies could have reported test-retest reliability (i.e., participants performed two scans), most did not mention it. Seven mentioned reliability in the discussion and only one reported test-retest reliability at the subject level; Siegle et al. (2012) reported “sgACC z scores and reactivity had moderate test-retest reliability in controls undergoing testing approximately 16 weeks apart (N=27; r=0.39

[P=0.04]). All but 1 had a pretest z score less than 0.5, and all but 2 had a posttest z score less than 0.5, suggesting stability within a restricted range." Other studies that mention reliability describe stability of group effects. For example, "Test-retest effects were accounted for by the healthy control group, who underwent the same scans at the same time points" (Walsh et al., 2007) is often reported in the discussion. This technique, while valuable, does not yield estimates of test-retest reliability at the individual subject level; the absence of a main effect of Time is evidence of the lack of a mean shift, but not of the stability of participants ranks.

1.3. rtfMRI-nf Studies Review

Interventions that use biological measures as real-time targets, including rtfMRI-nf also implicitly assume reliability. rtfMRI-nf trains patients to regulate the hemodynamic activity in regions of interest (Decharms, 2008) with the hope that changing a causal mechanism will result in symptom changes. rtfMRI-nf appears useful for several clinical populations, including patients with MDD (Thibault et al., 2018). Most patients can learn volitional control of hemodynamic activity in a targeted brain region (Fovet et al., 2015) which has been associated with clinical improvements (Fovet et al., 2015; Linden, 2014; Linden et al., 2012; Young et al., 2014) suggesting potential translational applications (Decharms, 2008; Ruiz et al., 2014; Thibault et al., 2018). An implicit assumption of rtfMRI-nf is that the signal measured on one day represents the same quantity measured on subsequent days, and thus performance on that metric can be trained over days. Consequently, test-retest reliability seems a strong prerequisite. Thus, as for prediction studies, we considered whether test-retest reliability is being reported in the fMRI neurofeedback literature.

1.3.1. Method

A PubMed search with the key words "(neurofeedback AND fMRI) OR rt-fMRI-nf) AND (depression OR MDD OR major depressive disorder" provided 44 results. After removing articles not including rtfMRI-nf or patients suffering from MDD, we were left with 11 studies (Table 2).

Table 2: rt-fMRI-nf studies in Major Depressive Disorder and mention and possibility of test-retest reliability

Reference	Neurofeedback	ROI	Mention of the reliability of the signal	Possibility to test signal reliability	How could they look at reliability
Linden et al. (2012)	Upregulation	Functional localizer of brain areas involved in the generation of positive emotions (e.g., VLPLC, insula)	No	Yes	Same regions selected by the localizer on different sessions
Zotev, et al., (2014)	Upregulation	Left amygdala (anatomical)	No	No	-
Young et al. (2014) [a]	Upregulation	Left amygdala (anatomical)	No	Yes	Reliability of fMRI signal in ROI
Yuan et al. (2014)[a]	Upregulation	Left amygdala (anatomical)	No	No	-
Zotev et al. (2016)[a]	Upregulation	Left amygdala (anatomical)	No	No	-
Hamilton et al. (2016)	Downregulation	Functional localizer of the salience network	No	No	-
Young et al. (2017)[b]	Upregulation	Left amygdala (anatomical)	No	Yes	Reliability of fMRI signal in ROI

Young, Misaki, et al. (2017)[b]	Upregulation	Left amygdala (anatomical)	No	Yes	Reliability of fMRI signal in ROI
Young et al. (2018)[b]	Upregulation	Left amygdala (anatomical)	No	Yes	Reliability of fMRI signal in ROI
MacDuffie et al. (2018)	Upregulation and downregulation	Functional localizer of ACC	No	No	-
Mehler et al. (2018)	Upregulation	Functional localizer of brain areas involved in seeing positive versus neutral pictures (e.g., insula and striatum)	No	Yes	Same regions selected by the localizer on different sessions

ACC: Anterior Cingulate Cortex; VLPFC: Ventrolateral Prefrontal Cortex

References associated with the same letter refer to the same data set

1.3.2. Results

None of the examined fMRI-nf studies reported on the reliability of the signal being trained (Table 2 and specific discussion of functional localizers in Box 4 in supplements).

1.4. Conclusions Thus Far

MDD studies using fMRI for clinical prediction or treatment rarely mention reliability, mirroring the more general fMRI literature (for meta-analysis, see Elliott et al., 2020). This lack of reporting could be due to failure to consider psychometrics important, or systematic decisions not to report observed low reliabilities. Indeed, reliability in published fMRI research in non-clinical studies, across protocols, tasks, regions of interest, psychological functions, and retest intervals have been fairly low (ICC~0.50),

with most published studies reporting values between 0.33-0.66. These values are mostly below "good" reliability thresholds for psychometrically sound clinical tests (~0.6).

2. POTENTIAL WAYS TO OPTIMIZE TEST-RETEST RELIABILITY IN fMRI/rtfMRI-NF

To facilitate reporting of reliability in clinical studies as part of every-day neuroimaging-science, the remainder of this article is dedicated to introducing ways to report, improve, and increase clinical applicability of test-retest reliability for fMRI in clinical populations. We apply and evaluate these suggestions in two published data sets (Siegle et al., 2012; Young et al., 2017b).

There is already a strong literature on optimizing preprocessing, which can increase measurement of true signal, and thus reliability (Andersson et al., 2001; Miki et al., 2000; Oakes et al., 2005; Zhilkin and Alexander, 2004). We therefore begin by considering whether using alternate ways of indexing task-related reactivity in single-subject data with optimized preprocessing lead to improved test-retest reliability.

As each combination of task, design, scanner, preprocessing and analysis strategy has a unique value of reliability that cannot necessarily be generalized to other studies (Braver et al., 2010), it may be useful to have standardized generally applicable methods to find out which regions and analysis methods have sufficient psychometric qualities to be used as biomarkers or in which the signal is stable enough to be able to give relevant feedback of its activation.

2.1. Optimize indices of task-related reactivity

The first possibility we consider involves optimizing indices for task related reactivity in fMRI. This Blood Oxygen Level Dependent (BOLD) response is generally considered to be convolution of the time-course of neural activity with a physiological hemodynamic response. Mis-specification of the shape of BOLD reactivity can introduce inefficiency and noise into estimates, which decreases reliability in human (Handwerker et al., 2012; Lindquist et al., 2009; Shan et al., 2014) and animal models (Peng et al., 2019). If, for example, neural responses to task stimuli are sustained in depression rather than increased in amplitude (e.g., Mandell, Siegle, Shutt, Feldmiller, & Thase, 2014), standard indices such as the amplitude of the canonical BOLD response may not capture relevant aspects of the pathology.

Thus, we propose evaluating indices such as the average amplitude, area under the curve and timing/shape of the curve of the BOLD response in addition to its canonical amplitude. Gamma variate models, in particular, yield parameters for onset, rise and fall slopes, and magnitude of hemodynamic responses (e.g., Larson et al., 2006), which can be evaluated for reliability. Similarly, including temporal and dispersion derivatives can account for individual differences in peak response timing and small differences in HRF length, providing larger test-retest reliability values (Fournier et al., 2014).

2.2. Examine Regions with Voxel-Wise High Test-Retest Reliability

When considering task-related reactivity in a region of interest (ROI), it is useful to reduce voxelwise reactivity to a single or few indices which capture reactivity across the region as a whole. The same consideration applies for reliability. Caceres et. al. (2009) suggest computing the ICC in each voxel within a region of interest (ROI) and reporting the median ICC as an index of region's test-retest reliability. This approach has been applied practically (Fournier et al., 2014; Lois et al., 2018). However, several potential biomarkers and neurofeedback targets identified in the literature, including the amygdala (Lebow and Chen, 2016; Young et al., 2014) and the sgACC (Siegle et al., 2012), consist of subregions with anatomical and functional heterogeneity (Hrybouski et al., 2016; Palomero-Gallagher et al., 2019). Their reliability may not be the same across these sub-divisions (Brabec et al., 2010; Janak and Tye, 2015; LeDoux, 2012). Therefore, it is possible that only some parts of ROIs may have adequate reliability and that the median reliability will not capture the most reliable parts of the signal. Just as questionnaires are traditionally constructed by eliminating unreliable items from an initial theoretically plausible set (Sheatsley, 1983), an index that inherits solely from the reliable voxels may increase psychometric properties of the preserved portions of regions.

Ten years ago, Bennet and Miller (2010) suggested that voxelwise reliability constitutes the most rigorous criteria of reliability since it implies that the level of activity in all voxels should remain consistent between scans. Although few studies have used this approach, we contend the available psychometric arguments weight in favor of voxel-wise computation of ICCs, restricting "reliable" ROIs to those regions in which all voxels have good or excellent reliability.

2.3. Optimize Models to Account for Individual and Clinical Features

Minimizing sources of non-interest that could vary between administrations increases the reliability of acquired data (Lin and Monica Way, 2014). Some fMRI noise sources such as differences in instrumentation, time of day, motion, etc. can be controlled, to some extent, via design. Tasks can be selected which have few practice effects and pre-baseline training can remove practice and strategy-development effects (Barch and Mathalon, 2011; Palmer et al., 2018). Choosing as-simple-as-possible tasks can minimize the impact of non-task cognitive processes. Standardizing instructions and training procedures helps to ensure participants understand the task before the first administration (Barch and Mathalon, 2011). Effects of other time-varying noise sources, such as thermal and physiological noise, are routinely minimized via preprocessing procedures (Krüger and Glover, 2001).

That said, if sources of variation across time, such as physiological or cognitive features, cannot be fully managed within design or processing, statistical methods for adjusting test-retest reliability estimates for them (Atri et al., 2011; Hsiao et al., 2011; Laenen et al., 2006) may be important to consider. Indeed, individual differences in state anxiety can account for amygdala activation (Calder et al., 2011) and habituation (Sladky et al., 2012), and, variation in rumination in depression is continuously associated with individual differences in amygdala, hippocampal, and prefrontal reactivity to emotional stimuli (Mandell et al., 2014; Siegle et al., 2002). Thus, true signal differences due to anxiety, mood or other symptoms between scans, especially if test-retest reliability is being evaluated in the context of possible treatment-related effects, might account for apparently unreliable neural responses, particularly to emotional stimuli. Thus, it may be useful to account for individuals' differences that could change across time statistically in estimating reliability, e.g., via the inclusion of clinical covariates.

2.4. Examine Reliability Within Relevant Tasks and Clinical Populations

Estimating reliability in healthy participants or symptomatic individuals who do not receive intervention may help separate effects of symptom change from practice effects. Yet, these approaches can introduce other confounds (e.g., if a task is reliable in patients but not controls or non-treatment seeking symptomatic individuals). The majority of studies have examined reliability of fMRI data in homogenous samples of healthy, often young, university students (Bennett and Miller, 2010; Lois et al., 2018).

Studies reviewed in Table 1 that discuss reliability in MDD generally restrict their discussion to whether there was a main effect of Time in healthy controls. Generally, BOLD response variability is greater in forms of between-subject responses than within (Aguirre et al., 1998). A limitation of ICC is that simultaneous inclusion of within and between subject variability causes estimators to be affected by sample composition. As groups might differ in the degree to which regional signals are reliable between measurements (Fournier et al., 2014), and because ICCs are proportional to between subject variability, heterogeneous samples can produce different ICCs even with the same degree of within-subject reliability of test-retest values. Using only healthy control participants may underrepresent true variability or over represent measurement errors in the population of interest, yielding inaccurate reliability estimates. Similarly, non-treatment seeking patients differ from treatment seeking patients on many variables that could affect test-retest reliability, such as symptomatology and comorbidity (Galbaud Du Fort et al., 1999).

Thus, testing reliability in the population of interest may provide more accurate estimates. We therefore recommend the use of representative samples to create a voxel-wise, population- and task-specific map of test-retest reliability. For example, if a task is to be used to distinguish symptomatic from healthy individuals, this method should be applied to a mixed population of healthy and symptomatic participants prior to the clinical application of the task. If the purpose is to distinguish respondents and non-respondents to a treatment, we recommend assessing reliability among treatment-seeking patients.

3. EVALUATION OF SUGGESTED OPTIMIZATIONS IN A PROGNOSTIC NEUROIMAGING TREATMENT OUTCOME DATASET

We have described several approaches that could be useful when examining and seeking to improve test-retest reliability in service of clinical translation including R1) optimizing BOLD signal parameterization, R2) using regions or voxels with stronger psychometric properties, R3) accounting for within-individual changes and R4) studying relevant tasks and populations for the intended application. In this section we demonstrate feasibility of these approaches and examine whether they are useful when applied to a published clinical fMRI dataset (Siegle et al., 2012). Our code for these analyses is freely available from https://github.com/PICANlab/Reliability_toolbox in the folder named “activation_task_reliability”.

3.1. Method

The sample consisted of participants described in Siegle et al. (2012) augmented by the addition of 8 patients who completed the same protocol after that paper was submitted, yielding 57 patients with major depressive disorder (MDD), and 35 healthy control participants (see supplement for details of this dataset and its relationship to Siegle et al 2012). Briefly, participants with MDD completed a slow event-related task during 3T fMRI in which they labeled the valence of emotional words (here, as in the published dataset, we analyzed only nominally negative words) before and after 12-16 weeks of Cognitive Therapy.

We computed reliability estimates within 4 ROIs which the literature suggests may function as biomarkers for treatment response including the amygdala (Arnone et al., 2012; Godlewska et al., 2012; Sheline et al., 2001), dorsolateral prefrontal cortex (DLPFC, Koenigs and Grafman, 2009), rostral anterior cingulate cortex (rACC, Hunter et al., 2013) and subgenual cingulate cortex (sgACC, Siegle et al., 2012b; Straub et al., 2015; Taylor et al., 2018) (our region-wise definitions are included in Box 1 in Supplement).

3.1.1. Optimize the BOLD Signal

The BOLD response to negative words was modeled within participants using 4 different methods including 1) amplitude of a canonically shaped BOLD signal (using AFNI's 3dDeconvolve with a narrow tent function ('BLOCK5(1,1)', Cox, 1996), 2) Area under the curve (via multiple regression of a delta function across 8 TRs using 3dDeconvolve, i.e. computed with Finite Impulse Response/FIR basis, with sum of betas as the parameter retained); 3) Peak amplitude from the same regressions as #2, and 4) a gamma variate model with parameters for onset-delay, rise-decay rate, and height. Voxelwise outliers outside the Tukey hinges were windsorized across participants and ICCs (3,1) were computed (Shrout and Fleiss, 1979) within individuals for each modeling method using custom Matlab code. While ICC(2,1) allows generalizing results obtained from different scanners, we chose to use ICC(3,1) to be able to compare with most of the literature, given that it is the most widely used ICC. This approach also allowed us to examine the importance of including scanner as a covariate in 3.1.3.

3.1.2. Compute Voxelwise Reliability

To measure the benefit of identifying reliable voxels, we calculated the mean, median and standard deviation of the ICCs in each of the ROIs for each modeling method and each group.

3.1.3. Include Clinical and Design Related Measures

We examined whether indices of reliability increased when clinical and design-related measures were included. As the ICC does not easily allow inclusion of covariates, we used semi partial correlations within the context of multiple regressions with and without covariates to assess changes in reliability, where covariates were pre and post clinical measures, as:

$$\widehat{Post} = \beta 0 + \beta(1 \rightarrow n)covariates + \beta(n+1)Pre$$

This model accounts for the potential that participants who show little change in symptoms may have better test-retest reliability. Modelling these clinical effects at the group level should make it possible to identify variance unique to test-retest reliability.

We included indices of pre- and post-treatment depressive symptomatology (Beck Depression Inventory; BDI, Beck et al., 1996), state and trait anxiety (Spielberger, 1983), rumination (Nolen-Hoeksema et al., 1993), and sleepiness (Johns, 1991) administered on the scan day, the scanner on which data were acquired, and participant's group when patients and controls were considered in one sample, coded as dummy variables, as covariates. Missing data were imputed via regression from the other administered measures also used as covariates.

A primary question was whether any of the proposed techniques described above, including different BOLD models, accounting for voxelwise variability, and the use of covariates, would differentially affect reliability estimates (i.e., semi-partial correlations). As such, after computing reliability estimates at each voxel, we rank ordered them across all permutations of BOLD estimate parameters (6 parameters) and the use or non-use of covariates (2 conditions) at each voxel per ROI, yielding 12 x #-voxels rankings per ROI. Following a Kolgomorov-Smirnov test justifying the need to use non-parametric tests, we report a Kruskal-Wallis test to determine whether the rankings differed across models in each ROI. If they did, as a simple effects test, we generated confidence intervals around the mean of rankings for each of the 12 conditions via a one-way ANOVA (via Matlab's multcompare function). Non-overlapping confidence intervals are interpretable as significant differences between one condition and any other. To display them we generated figures showing the mean

of rankings for each condition, which will be numbers on the order of 1 to 12 x # voxels, with higher means representing being at the top of the rankings across many voxels within the ROI.

3.1.4. Use Clinically Representative Samples

All analyses were conducted on the whole sample (controls and patients) to establish likely reliability of tests that could be used to discriminate groups, and on patients only, to establish likely reliability of clinical prognostic and change indicators. We considered multiple reliability effect size thresholds which might be used in other studies (0.4 and 0.6 for fair and good reliability and 0.7, and 0.75 for traditional labels of the data as “reliable” and clinically meaningful).

3.1.5. Type 1 error control

As 1) each of the hypotheses and regions examined for this manuscript was considered a different family of tests and 2) we want our results to generalize to reliability as it is reported in the confirmatory biomarker and neurofeedback literatures where only one region is generally examined, consistent with the literature on test-retest reliability in neuroimaging, type I error was not controlled across regions and hypotheses for ROI-wise statistics. For simple-effects tests of differences in rankings across conditions, we controlled for the number of conditions with a Bonferroni test. For voxelwise statistics we subjected all voxelwise residual maps to empirical cluster thresholding (AFNI’s 3dFWHMx and 3dClustSim, acf model with small-volume corrections for examined regions) using a p threshold (-pthr) based on each considered effect size threshold (see in supplementary materials, table S3 for more details).

3.2. Results and discussion

3.2.1. Optimizing the BOLD signal

ICC’s were uniformly low (<.3) for all BOLD parameterizations when entire ROIs were considered (Table 4). Kruskal Wallis tests did suggest differential reliability across our parameterizations (Table 5a). This held when the two outlying uniformly low reliability parameterizations (rise decay with and without covariates) were removed from consideration (Table 5b). Yet, there were non-overlapping confidence intervals among counts of rank orderings of parameterizations for voxelwise tests, suggesting that at least for some subsets of regions, some parameterizations were superior (Supplement Figure S1, Table S1). For example, in the full sample, for the amygdala, amplitude without covariates was superior to other parameters. Over all ROIs, the most

reliable parameters were amplitude, canonical amplitude, and height (Figure 1 shows voxelwise variation within a Priori ROIs for the height parameter) for the whole sample and amplitude, area under the curve, and height for only patients (Figure S1 and Table S1). However, looking at ROIs and samples independently, the parameter offering the highest levels of reliability varied.

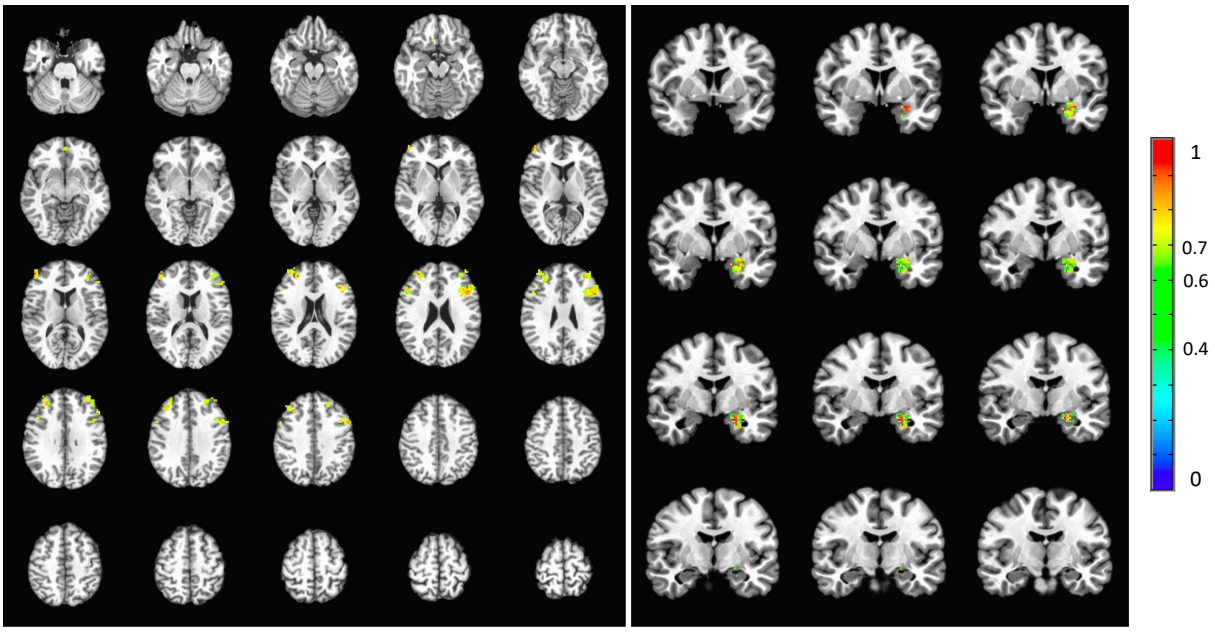

Figure 1: Test-retest reliability in ROIs estimated with voxel wise ICCs using height parameter, a threshold of ICC>0.4 and cluster correction applied for this threshold in A. Siegle et al. (2012) dataset of patients and B. Young et al. (2017) data set of the transfer run in the experimental group (signal with training) preprocessed with the TBV style pipeline.

3.2.2. Voxelwise reliability

In the whole sample, moderate reliability (ICC>.4) in clusters large enough to infer significance was observed in the DLPFC using the canonical amplitude model and in the amygdala using amplitude (Table 3). “Good” (ICC>.6) reliability was reached in clusters large enough to infer significance when only the patients were considered, using amplitude and height in the DLPFC. These levels of voxelwise test-retest reliability were higher than using the median or mean value of ICCs within whole ROIs (Table 4). Levels of generally accepted reliability for clinical measures (ICC>.7) were not observed in clusters large enough to report.

Table 3: Table of number of voxels reaching different reliability thresholds for each sample, first level parameter, and ROI with cluster correction applied.

	ROI	Amydgala (242 voxels)		DLPFC (2675 voxels)		rACC (865 voxels)		sgACC liberally thresholded (33 voxels)		sgACC conservatively thresholded (18 voxels)	
Population	Reactivity model	ICC thresholds		ICC thresholds		ICC thresholds		ICC thresholds		ICC thresholds	
		0.4	0.6	0.4	0.6	0.4	0.6	0.4	0.6	0.4	0.6
Controls & patients	Canonical amplitude	0	0	465	0	0	0	0	0	0	0
	Amplitude	66	0	5	0	0	0	0	0	0	0
	Area under the curve	10	0	0	0	0	0	0	0	0	0
	Onset delay	0	0	0	0	0	0	0	0	0	0
	Rise decay	0	0	0	0	0	0	0	0	0	0
	Height	0	0	290	2	0	0	0	0	0	0
Patients	Canonical amplitude	0	0	299	6	6	0	0	0	0	0
	Amplitude	24	0	0	0	0	0	0	0	0	0
	Area under the curve	0	0	0	0	0	0	0	0	0	0
	Onset delay	0	0	0	0	0	0	0	0	0	0
	Rise decay	0	0	0	0	0	0	0	0	0	0
	Height	0	0	374	5	5	0	2	0	1	0

Table 4: Table of mean, standard deviation and median values of ICCs for each sample, reactivity model, and ROI.

Population	Reactivity model	Amygdala	DLPFC	rACC	sgACC liberally thresholded	sgACC conservatively thresholded
Controls & patients	Canonical amplitude	0.11 (±0.09); 0.11	0.24 (±0.16);	0.09 (±0.10);	0.15 (±0.08); 0.13	0.17 (±0.09); 0.18

			0.26	0.09		
	Amplitude	0.23(±0.14); 0.22	0.12 (±0.11); 0.12	0.11 (±0.10); 0.12	-0.01 (±0.13); -0.04	-0.04 (±0.14); -0.08
	Area under the curve	0.13 (±0.14); 0.12	0.08 (±0.10); 0.07	0.03 (±0.11); 0.03	-0.03 (±0.09); -0.04	-0.06 (±0.10); -0.07
	Onset delay	0 (±0.09); -0.01	0.01 (±0.09); 0	0 (±0.10); 0	0 (±0.08); 0.01	-0.01 (±0.10); 0
	Rise decay	0 (±0); 0	0 (±0); 0	0 (±0); 0	0 (±0); 0	0 (±0); 0
	Height	0.08 (±0.10); 0.09	0.21 (±0.15); 0.23	0.13 (±0.12); 0.14	0.16 (±0.12); 0.17	0.18 (±0.12); 0.23
Patients	**Canonical amplitude**	0.09(±0.11); 0.11	0.22 (±0.16); 0.23	0.08 (±0.14); 0.08	0.10 (±0.12); 0.07	0.14 (±0.15); 0.12
	Amplitude	0.22 (±0.15); 0.22	0.11 (±0.13); 0.11	0.10 (±0.13); 0.11	-0.06 (±0.15); -0.07	-0.08 (±0.14); -0.08
	Area under the curve	0.13(±0.14); 0.12	0.6 (±0.12); 0.03	0.03 (±0.13); 0.04	-0.08 (±0.13); -0.08	-0.10 (±0.13); -0.09
	Onset delay	-0.01 (±0.12); -0.01	0.01 (±0.12); 0	-0.01 (±0.13); -0.01	0.02 (±0.11); 0.02	0.01 (±0.12); 0.05
	Rise decay	0 (±0); 0	0 (±0); 0	0 (±0); 0	0 (±0); 0	0 (±0); 0
	Height	0.09 (±0.12); 0.08	0.22 (±0.16); 0.23	0.12 (±0.15); 0.13	0.16 (±0.17); 0.18	0.17 (±0.17); 0.21

Mean (±standard deviation); median

Table 5a: Table of Kruskal Wallis tests' output for each sample, reactivity model with and without covariates, and ROI with Bonferroni correction applied.

Population	Amygdala	DLPFC	rACC	sgACC liberally thresholded	sgACC conservatively thresholded
Controls & patients	**H(11)=1414.67, p<0.001**	**H(11)=12717.07, p<0.001**	**H(11)=4794.14, p<0.001**	**H(11)=206.47, p<0.001**	**H(11)=118.32, p<0.001**

Patients	**H(11)=1233.13, p<0.001**	**H(11)=10371.75, p<0.001**	**H(11)=4477.55, p<0.001**	**H(11)=240.89, p<0.001**	**H(11)=136.93, p<0.001**

Note: Applying Bonferroni correction for 6 reactivity models with and without covariates (p<0.05/12=0.004).

Table 5b: Table of Kruskal Wallis tests' output for each sample, reactivity model with and without covariates, and ROI with Bonferroni correction applied, without rise decay.

Population	Amygdala	DLPFC	rACC	sgACC liberally thresholded	sgACC conservatively thresholded
Controls & patients	**H(9)=285.88, p<0.001**	**H(9)=4876.99, p<0.001**	**H(9)=644.19, p<0.001**	**H(9)=58.90, p<0.001**	**H(9)=40.55, p<0.001**
Patients	**H(9)=25.68, p=0.002**	**H(9)=1588.15, p<0.001**	**H(9)=190.42, p<0.001**	**H(9)=108.21, p<0.001**	**H(9)=67.20, p<0.001**

Note: Applying Bonferroni correction for 6 reactivity models with and without covariates (p<0.05/10=0.005).

3.2.3. Clinical and Design Related Measures

The addition of covariates never resulted in significantly higher average ranks for semi partial correlations in any ROI, in the whole sample or just the patients (Figure S1). In other words, adding covariates did not improve the reliability, and in some instances made it worse.

4. EVALUATION OF SUGGESTED OPTIMIZATIONS IN AN EMPIRICAL NEUROFEEDBACK DATASET

To further support the feasibility of applying these recommendations and to evaluate the consistency of their performance in a second dataset, we consider a published fMRI neurofeedback dataset (Young, Siegle, et al., 2017, code available from https://github.com/PICANlab/Reliability_toolbox in the folder named "rtfMRI-nf_reliability").

4.1. Method

This dataset constituted 18 patients in the experimental group who received amygdala neurofeedback and 16 patients in the control group who received parietal

neurofeedback. Briefly, participants completed two training scans on different days within 2 weeks, each including a "baseline" and "transfer" runs during which no feedback was presented. The analyzed task was a 40-second per block design during which participants alternately rested, worked to upregulate a target region during recall of positive memories, and did a distraction (counting) task (see supplement Box 5 for details of this dataset). Here, we focus on a) the baseline data on the two training days in control-feedback participants during recall of positive autobiographical memories prior to neurofeedback training. As their amygdala signal did not change over the course of the study at the group level (Young et al., 2017b), this allows us to examine test-retest reliability of the left amygdala signal without the influence of neurofeedback. b) the left amygdala signal during the two transfer runs in the experimental group, as this represents the effect of neurofeedback training. Activity during the two post-training transfer runs did not differ at the group level, allowing us to examine the test-retest reliability of the amygdala signal after neurofeedback training. Because this dataset only included patients with MDD, only the first 3 principles (i.e., optimization of the BOLD signal, computation of voxelwise reliability, and inclusion of clinical and design related measures) are evaluated in this dataset.

Feedback signal

To analyze the feedback signal averaged over the left amygdala we used the output of the script used in Young, Siegle, et al. (2017) that allowed computation of the feedback signal in real-time before considering the voxel-wise signal.

Voxel-wise

As rtfMRI-nf involves real-time preprocessing of the data, we sought to examine whether this kind of preprocessing could affect the test-retest reliability of the signal. We therefore performed data preprocessing emulating the real-time data processing performed by the commercially available neurofeedback software Turbo BrainVoyager (BrainVoyager, The Netherlands; henceforth "TBV style") and a more classic contemporary post-hoc preprocessing stream (here referred to as "standard preprocessing"). Both streams were implemented using AFNI.

- <u>TBV style preprocessing</u>

Turbo BrainVoyager performs the following functions in real-time: 3D motion correction, spatial smoothing, and drift removal via the design matrix. We used AFNI to approximate these steps. After spatially transforming the anatomical then functionals to the International Consortium for Brain Mapping 152 template, we then rescaled them

to conform to the Talairach atlas dimensions and then performed motion correction to the first image, spatial smoothing 4mm FWHM smoothing kernel and fourth order detrend for drift removal.

- Standard preprocessing

MRI pre-processing included despiking, volume registration and slice timing correction for all EPI volumes in a given exam. After applying an intensity uniformity correction on the anatomical, the anatomical was spatially transformed to the International Consortium for Brain Mapping 152 template and rescaled to conform to the Talairach atlas dimensions. Then, the fMRI data for each run were warped nonlinearly and the same spatial transformations were applied. The fMRI run was spatially smoothed within the grey matter mask using a Gaussian kernel with full width at half maximum (FWHM) of 4 mm. A first standard GLM analysis was then applied separately for each of the fMRI runs. The following regressors were included in the GLM model: six motion parameters and their derivatives as nuisance covariates to take into account possible artifacts caused by head motion, white matter and cerebrospinal fluid signals, and five polynomial terms for modeling drift.

4.1.1. Optimize the BOLD Signal

4.1.1.1. Amygdala signal

From each participant's real-time left amygdala signal we calculated an "amygdala signal" for each positive recall block minus the mean of the preceding rest block from the output of previously used scripts for real-time preprocessing (Young et al., 2017b), and recreated the feedback signal by taking the amount of activation at every TR during the experimental condition minus the mean activation in the previous rest condition, on the baseline run of control participants at visits 1 and 2 (signal without training) and on the transfer run of experimental participants at visits 1 and 2 (signal with training), independently. We then averaged the time course of the feedback signal over all happy blocks. We summarized the activation for each participant for each visit by either a mean of the amygdala signal or by fitting the time course with a gamma variate model with parameters for onset-delay, rise-decay-rate, and height (see Methodological choice to fit gamma variates in Supplement Box 2 for more information of this methodological choice).

ICC(3,1) estimates were computed (Shrout and Fleiss, 1979) independently on the estimates of the feedback signal with and without training.

4.1.1.2. Voxelwise signal

The same reactivity models as in the treatment outcome dataset were applied (see part 3.2.1.1) to data preprocessed with both types of preprocessing but adapted to this design (AFNI tent parameters to accommodate 40 s blocks as BLOCK(40,1), and area under the curve across entire blocks).

4.1.2. Compute Voxelwise Reliability

As in the treatment outcome data set, to measure the benefit of identifying reliable voxels, we calculated the mean, median and standard deviation of the ICCs in the left amygdala for each model, group, and additionally for both preprocessing pipelines.

4.1.3. Include Clinical and Design Related Measures

As in the treatment outcome data set, semi partial correlations were computed with and without covariates. We included indices of depressive symptomatology (Beck Depression Inventory; BDI, Beck et al., 1996), state and trait anxiety (Spielberger, 1983), sleepiness and drowsiness administered on the scan day, and the scanner on which data were acquired coded as dummy variables, as covariates. There was no missing data. We then compared the semi-partial correlations across all models of individual responses with and without covariates for each group and preprocessing pipeline as in section 3.1.3, to understand which models offered adequate test-retest reliability and whether there were differences between them.

4.1.4. Type 1 error control

As discussed in section 3.1.5, cluster correction was applied on voxelwise statistics (further details in supplement table S4).

4.2. Results and discussion

4.2.1. Optimizing the BOLD signal

4.2.1.1. Amygdala signal

The mean amygdala signals with and without training showed poor reliability (ICCs<0.1). When the amygdala signal within the left amygdala was fit using a gamma variate function, the onset-delay and height parameters showed fair reliability for the signal without training (ICC=0.54 and ICC=0.47, respectively), with all other models, including those with training, showing minimal reliability (ICC<.1). Therefore, it appears that the shape of the signal without training is consistent across sessions and that the signal in the left amygdala is more reliable when unchanged by training, which is consistent with the assumption that training is changing the signal over time.

4.2.1.2. Voxel-wise signal

Kruskal Wallis tests suggested there were differences between the parameters in reliability (Tables 6a and 6b). In particular, reliability for the height parameter (as well as amplitude for the signal without training) was higher than for other parameters (Figure S1). The height parameter also yielded a large enough cluster to infer significance for "excellent" (ICC>.7) reliability in both samples (Table 7, Figure 1 for illustration).

The use of the standard preprocessing stream had non-significantly-different reliabilities from the stream emulating the real-time preprocessing run by Turbo BrainVoyager over all parameters with or without covariates, with the exception of the height parameter without covariates, which showed higher reliability with TBV style preprocessing than with standard preprocessing in the signal without training (see Figure S1).

Table 6a: Table of Kruskal Wallis tests' output for each sample, reactivity model with and without covariates in the left amygdala with Bonferroni correction applied.

Population	Amygdala
Without training - Control - Baseline	H(23)=2964.56, p<0.001
With training - Experimental - Transfer	H(23)=3142.17, p<0.001

Note: Applying Bonferroni correction for 6 reactivity models with and without covariates (p<0.05/12=0.004).

Table 6b: Table of Kruskal Wallis tests' output for each sample, reactivity model with and without covariates in the left amygdala with Bonferroni correction applied, without rise decay.

Population	Amygdala
Without training - Control - Baseline	H(19)=1397.84, p<0.001
With training - Experimental - Transfer	H(19)=1702.57, p<0.001

Note: Applying Bonferroni correction for 6 reactivity models with and without covariates (p<0.05/10=0.005).

Table 7: Table of number of voxels reaching different reliability thresholds for each sample, preprocessing, and first level parameter with cluster correction applied.

ROI		Amygdala (214 voxels)							
Preprocessing		BV style				Standard			
Population	First level model	ICC thresholds				ICC thresholds			
		0.4	0.6	0.7	0.75	0.4	0.6	0.7	0.75
Without training - Control - Baseline	Canonical amplitude	0	0	0	0	0	0	0	0
	Amplitude	52	16	6	2	35	0	0	0
	Area under the curve	0	0	0	0	40	0	0	0
	Onset-delay	0	0	0	0	0	0	0	0
	Rise-decay	0	0	0	0	0	0	0	0
	Height	78	26	13	13	53	24	9	5
With training - Experimental - Transfer	Canonical amplitude	0	0	0	0	0	0	0	0
	Amplitude	66	4	2	2	42	11	3	2
	Area under the curve	0	0	0	0	0	0	0	0
	Onset-delay	0	4	4	4	0	5	5	5
	Rise-decay	0	0	0	0	0	0	0	0
	Height	159	81	25	16	73	47	24	21

4.2.2. Voxelwise reliability

Some voxelwise ICC values obtained were higher than those computed on the real-time signal covering the entire left amygdala or mean or median ICC values computed over the entire left amygdala (Table 5 vs statistics reported in 4.2.1.1 and Table 6), with some clusters achieving an excellent level of reliability (ICC>.7, see Table 5) for standard and TBV-like preprocessing both for the trained and untrained signals, which did not occur for the region as a whole.

Table 8: Table of mean, standard deviation and median values of ICCs for each sample, preprocessing, and first level parameter with cluster correction applied.

Preprocessing	TBV style	Standard

Without training - Control - Baseline	**Canonical amplitude**	-0.07 (±0.21); -0.09	0.01 (±0.24); 0
	Amplitude	0.29 (±0.2); 0.3	0.26 (±0.22); 0.27
	Area under the curve	0.02(±0.21); 0.01	0.21 (±0.23); 0.18
	Onset-delay	-0.03 (±0.23); -0.05	-0.11 (±0.20); -0.14
	Rise-decay	NA (±NA); NA	NA (±NA); NA
	Height	0.36 (±0.23); 0.33	0.17 (±0.38); 0.24
With training - Experimental - Transfer	**Canonical amplitude**	-0.11 (±0.21); -0.12	0.08 (±0.21); 0.09
	Amplitude	0.3 (±0.18); 0.31	0.26 (±0.21); 0.25
	Area under the curve	0.06 (±0.20); 0.07	0.13 (±0.18); 0.13
	Onset-delay	0.02 (±0.24); -0.02	-0.05 (±0.24); -0.13
	Rise-decay	NA (±NA); NA	NA (±NA); NA
	Height	0.52 (±0.19); 0.56	0.35 (±0.28); 0.34

Mean (±standard deviation); median

4.2.3. Clinical and Design Related Measures

4.2.3.1. Amygdala signal

Adding covariates when computing semi-partial correlations over the mean amygdala signal improved descriptively reliability estimates for the signal without training (from sr=0.11 to sr=0.14) as well as parameters tested to fit the signal with training (mean: from sr=0.06 to sr=0.12, onset-delay: from sr=0.14 to sr=0.21, rise-decay: from sr=0.03 to sr=0.14, height: from sr=0.16 to sr=0.29) although in no case we did achieve a fair level of reliability (sr<0.4).

4.2.3.2. Voxelwise signal

The addition of covariates in never resulted in higher average ranks of semipartial correlation distributions on the untrained or trained signal preprocessed with the TBV-like or standard pipeline (Figure S1).

5. GENERAL DISCUSSION

As stated in a recent meta-analysis (Elliott et al., 2020), task fMRI reliability is not systematically evaluated and when it is, task-related fMRI measures show poor reliability. Our literature review shows that both prognostic and interventional fMRI studies in MDD, which might otherwise be poised for clinical translation, also do not attend to reliability. We demonstrate that by attending to some fairly simple principles, we can achieve fair to good reliability in a clinical prediction outcome dataset and excellent reliability in a neurofeedback fMRI study dataset (Figure 1). These principles include careful modeling of the BOLD signal, identification of reliable voxels within regions of interest, and calculation of reliability in the population for which translational applications are being considered. Across both datasets, the height parameter from a gamma variate function was the most reliable way to model the BOLD signal, especially among patients with MDD, in some regions of interest, and was, in some combinations of region and population or training condition, more reliable than canonical amplitude, though in other cases the reverse was true (Table 3 and 5 and Figure S1). Consequently, we recommend that researchers explore multiple ways of modeling the BOLD signal, particularly including gamma variate modeling in MDD, before concluding their experiment has low reliability. It may also be helpful for software for real-time analysis of fMRI data to implement alternative, potentially more reliable ways of characterizing BOLD responses in real-time.

Increasingly, functional differentiation of sub-regions of subcortical structures such as the amygdala has been acknowledged as important for fMRI (Balderston, Schultz, Hopkins, & Helmstetter, 2014; Ball et al., 2007; Michely, Rigoli, Rutledge, Hauser, & Dolan, 2020; Roy et al., 2009). The comparison of test-retest reliability estimates obtained on the feedback signal averaged over the whole amygdala versus these same estimates computed voxelwise in the neurofeedback dataset suggest non-uniformity across the amygdala in signal reliability as well; the extent to which these differences explain previous results localizing function to subregions is unclear. Thus, we suggest it may be useful to use a voxel-wise or subregion approach to estimating test-retest reliability. Indeed, this method reveals significantly large clusters of voxels with excellent test-retest reliability in the left amygdala which could be used as masks for neurofeedback targets; our method is easily feasible for new studies. Such excellent reliability, which is a prerequisite for clinical translation, was not attained in

our dataset, using the more common computation of median ICCs for each ROI (e.g., as recommended by Caceres et al., 2009) (see Tables 4 and 6).

Contrary to our hypotheses, we did not find that adding covariates to the model, including the scanner on which participants were run and severity, which did change as a function of intervention, improved test-retest reliability in these datasets (Figure S1) in ROI-based or whole-brain analyses (see Figure S2). That said, covariates may still be useful to include in other datasets – we recommend exploring this option further before dismissing their utility.

Reliability did vary by whether the entire sample or only patient's data were included and by whether or not participants were trained on the task, supporting the potential utility of quantifying reliability on tasks and populations that are relevant for the clinical application intended (Tables 3 and 5 and Figure S1).

There are several limitations of this review and analyses. As we have focused only on MDD, it is unclear whether our conclusions apply transdiagnositically. Improving reliability may require different strategies in other diseases, such as Parkinsons, due to age-related atrophy, increased movement, and differences in neurovascular coupling (Lecrux et al., 2019; Paek et al., 2019). There are many fMRI-based metrics we could have examined, including functional connectivity, volumetric measures, and resting state designs, which all provoke unique considerations for optimizing test-retest reliability, some of which have been explored elsewhere (e.g., Noble et al., 2019). Here, we focused on regional BOLD activity as it is a common feature of prediction and neurofeedback studies. Our published data sets had relatively small number of subjects. This is typical for most clinical fMRI studies, but does raise the concern that the sample is too small and underpowered. Therefore, we strongly encourage the replication of these results and that is also why we have applied these suggestions to two different data sets.

6. CONCLUSIONS

To summarize, demonstrating that mechanistic indices are reliable is important before their clinical adoption in prediction or treatment-development. The literature in these areas has implicitly accepted this assumption without testing it. Other non-clinical fMRI studies have shown many of the regions targeted in clinical fMRI studies have fairly low test-retest reliability, which was largely replicated using the most common analytic techniques in our datasets. Yet, we have suggested a few principles that

appear to improve the test-retest reliability of the obtained mechanistic signals, have shown their feasibility in two previously published fMRI data sets, and have made code publicly available so that researchers with minimal mathematical and programming knowledge can implement them. Wider adoption of these methods could help to realize the potential of clinical fMRI and could extend to improving psychometrics for other time-varying mechanistic indices.

Acknowledgements

This work was supported by the National Institutes of Health/ National Institute of Mental Health [grant numbers MH115927, MH106591, MH074807, MH58356, MH69618] and through the Pittsburgh Foundation Emmerling Fund [grant number M2007-0114]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Summary declaration of interest

Declarations of interest: none.

REFERENCES

Aguirre, G.K., Zarahn, E., D'esposito, M., 1998. The variability of human, BOLD hemodynamic responses. Neuroimage. https://doi.org/10.1006/nimg.1998.0369

Anand, A., Li, Y., Wang, Y., Gardner, K., Lowe, M.J., 2007. Reciprocal Effects of Antidepressant Treatment on Activity and Connectivity of the Mood Regulating Circuit: An fMRI Study. J. Neuropsychiatry Clin. Neurosci. https://doi.org/10.1176/jnp.2007.19.3.274

Andersson, J.L.R., Hutton, C., Ashburner, J., Turner, R., Friston, K., 2001. Modeling geometric deformations in EPI time series. Neuroimage. https://doi.org/10.1006/nimg.2001.0746

Arnone, D., 2019. Functional MRI findings, pharmacological treatment in major depression and clinical response. Prog. Neuro-Psychopharmacology Biol. Psychiatry. https://doi.org/10.1016/j.pnpbp.2018.08.004

Arnone, D., McKie, S., Elliott, R., Thomas, E.J., Downey, D., Juhasz, G., Williams, S.R., Deakin, J.F.W., Anderson, I.M., 2012. Increased amygdala responses to sad but not fearful faces in major depression: Relation to mood state and pharmacological treatment. Am. J. Psychiatry. https://doi.org/10.1176/appi.ajp.2012.11121774

Atri, A., O'Brien, J.L., Sreenivasan, A., Rastegar, S., Salisbury, S., DeLuca, A.N., O'Keefe, K.M., LaViolette, P.S., Rentz, D.M., Locascio, J.J., Sperling, R.A., 2011. Test-retest reliability of memory task functional magnetic resonance imaging in alzheimer disease clinical trials. Arch. Neurol. https://doi.org/10.1001/archneurol.2011.94

Balderston, N.L., Schultz, D.H., Hopkins, L., Helmstetter, F.J., 2014. Functionally distinct amygdala subregions identified using DTI and high-resolution fMRI. Soc. Cogn. Affect. Neurosci. 10, 1615–1622. https://doi.org/10.1093/scan/nsv055

Ball, T., Rahm, B., Eickhoff, S.B., Schulze-Bonhage, A., Speck, O., Mutschler, I., 2007. Response Properties of Human Amygdala Subregions: Evidence Based on Functional MRI Combined with Probabilistic Anatomical Maps. PLoS One 2, e307. https://doi.org/10.1371/journal.pone.0000307

Barch, D.M., Mathalon, D.H., 2011. Using brain imaging measures in studies of procognitive pharmacologic agents in schizophrenia: Psychometric and quality assurance considerations. Biol. Psychiatry.

https://doi.org/10.1016/j.biopsych.2011.01.004

Bartko, J.J., 1976. On various intraclass correlation reliability coefficients. Psychol. Bull. https://doi.org/10.1037/0033-2909.83.5.762

Bartko, J.J., 1966. The Intraclass Correlation Coefficient as a Measure of Reliability. Psychol. Rep. https://doi.org/10.2466/pr0.1966.19.1.3

Beck, A., Steer, R., Brown, G., 1996. Beck Depression Inventory II manual (2nd ed. Ed.), … for Beck Depression Inventory-II.

Benedetti, F., Radaelli, D., Bernasconi, A., Dallaspezia, S., Colombo, C., Smeraldi, E., 2009. Changes in medial prefrontal cortex neural responses parallel successful antidepressant combination of venlafaxine and light therapy. Arch. Ital. Biol.

Bennett, C.M., Miller, M.B., 2013. FMRI reliability: Influences of task and experimental design. Cogn. Affect. Behav. Neurosci. https://doi.org/10.3758/s13415-013-0195-1

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. https://doi.org/10.1111/j.1749-6632.2010.05446.x

Brabec, J., Rulseh, A., Hoyt, B., Vizek, M., Horinek, D., Hort, J., Petrovicky, P., 2010. Volumetry of the human amygdala - An anatomical study. Psychiatry Res. - Neuroimaging. https://doi.org/10.1016/j.pscychresns.2009.11.005

Braver, T.S., Cole, M.W., Yarkoni, T., 2010. Vive les differences! Individual variation in neural mechanisms of executive control. Curr. Opin. Neurobiol. https://doi.org/10.1016/j.conb.2010.03.002

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage. https://doi.org/10.1016/j.neuroimage.2008.12.035

Calder, A.J., Ewbank, M., Passamonti, L., 2011. Personality influences the neural responses to viewing facial expressions of emotion. Philos. Trans. R. Soc. B Biol. Sci. https://doi.org/10.1098/rstb.2010.0362

Canli, T., Cooney, R.E., Goldin, P., Shah, M., Sivers, H., Thomason, M.E., Whitfield-Gabrieli, S., Gabriels, J.D.E., Gotlib, I.H., 2005. Amygdala reactivity to emotional faces predicts improvement in major depression. Neuroreport. https://doi.org/10.1097/01.wnr.0000174407.09515.cc

Chen, C.H., Ridler, K., Suckling, J., Williams, S., Fu, C.H.Y., Merlo-Pich, E.,

Bullmore, E., 2007. Brain Imaging Correlates of Depressive Symptom Severity and Predictors of Symptom Improvement After Antidepressant Treatment. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2006.09.018

Cicchetti, D. V., 1994. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. Psychol. Assess. https://doi.org/10.1037/1040-3590.6.4.284

Costafreda, S.G., Khanna, A., Mourao-Miranda, J., Fu, C.H.Y., 2009. Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. Neuroreport. https://doi.org/10.1097/WNR.0b013e3283294159

Cox, R.W., 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. https://doi.org/10.1006/cbmr.1996.0014

Cullen, K.R., Klimes-Dougan, B., Vu, D.P., Westlund Schreiner, M., Mueller, B.A., Eberly, L.E., Camchong, J., Westervelt, A., Lim, K.O., 2016. Neural Correlates of Antidepressant Treatment Response in Adolescents with Major Depressive Disorder. J. Child Adolesc. Psychopharmacol. https://doi.org/10.1089/cap.2015.0232

Davidson, R.J., Irwin, W., Anderle, M.J., Kalin, N.H., 2003. The neural substrates of affective processing in depressed patients treated with venlafaxine. Am. J. Psychiatry. https://doi.org/10.1176/appi.ajp.160.1.64

Decharms, R.C., 2008. Applications of real-time fMRI. Nat. Rev. Neurosci. https://doi.org/10.1038/nrn2414

Delaveau, P., Jabourian, M., Lemogne, C., Allaïli, N., Choucha, W., Girault, N., Lehericy, S., Laredo, J., Fossati, P., 2016. Antidepressant short-term and long-term brain effects during self-referential processing in major depression. Psychiatry Res. - Neuroimaging. https://doi.org/10.1016/j.pscychresns.2015.11.007

Dichter, G.S., Felder, J.N., Smoski, M.J., 2010. The effects of Brief Behavioral Activation Therapy for Depression on cognitive control in affective contexts: An fMRI investigation. J. Affect. Disord. https://doi.org/10.1016/j.jad.2010.03.022

Doerig, N., Krieger, T., Altenstein, D., Schlumpf, Y., Spinelli, S., Späti, J., Brakowski, J., Quednow, B.B., Seifritz, E., Holtforth, M.G., 2016. Amygdala response to self-critical stimuli and symptom improvement in psychotherapy for depression. Br. J. Psychiatry. https://doi.org/10.1192/bjp.bp.114.149971

869 Elliott, M., Knodt, A., Ireland, D., Morris, M., Poulton, R., Ramrakha, S., Sison, M.,
870 Moffitt, T., Caspi, A., Hariri, A., 2020. What is the Test-Retest Reliability of
871 Common Task-fMRI Measures? New Empirical Evidence and a Meta-Analysis.
872 Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2020.02.356
873 Fales, C.L., Barch, D.M., Rundle, M.M., Mintun, M.A., Mathews, J., Snyder, A.Z.,
874 Sheline, Y.I., 2009. Antidepressant treatment normalizes hypoactivity in
875 dorsolateral prefrontal cortex during emotional interference processing in major
876 depression. J. Affect. Disord. https://doi.org/10.1016/j.jad.2008.04.027
877 Fang, J., Egorova, N., Rong, P., Liu, J., Hong, Y., Fan, Y., Wang, X., Wang, H., Yu,
878 Y., Ma, Y., Xu, C., Li, S., Zhao, J., Luo, M., Zhu, B., Kong, J., 2017. Early cortical
879 biomarkers of longitudinal transcutaneous vagus nerve stimulation treatment
880 success in depression. NeuroImage Clin.
881 https://doi.org/10.1016/j.nicl.2016.12.016
882 Fitzgerald, P.B., Sritharan, A., Daskalakis, Z.J., De Castella, A.R., Kulkarni, J., Egan,
883 G., 2007. A functional magnetic resonance imaging study of the effects of low
884 frequency right prefrontal transcranial magnetic stimulation in depression. J.
885 Clin. Psychopharmacol. https://doi.org/10.1097/jcp.0b013e318151521c
886 Fonseka, T.M., MacQueen, G.M., Kennedy, S.H., 2018. Neuroimaging biomarkers as
887 predictors of treatment outcome in Major Depressive Disorder. J. Affect. Disord.
888 https://doi.org/10.1016/j.jad.2017.10.049
889 Forbes, E.E., Olino, T.M., Ryan, N.D., Birmaher, B., Axelson, D., Moyles, D.L., Dahl,
890 R.E., 2010. Reward-related brain function as a predictor of treatment response
891 in adolescents with major depressive disorder. Cogn. Affect. Behav. Neurosci.
892 https://doi.org/10.3758/CABN.10.1.107
893 Fournier, J.C., Chase, H.W., Almeida, J., Phillips, M.L., 2014. Model specification
894 and the reliability of fMRI results: Implications for longitudinal neuroimaging
895 studies in psychiatry. PLoS One. https://doi.org/10.1371/journal.pone.0105169
896 Fovet, T., Jardri, R., Linden, D., 2015. Current Issues in the Use of fMRI-Based
897 Neurofeedback to Relieve Psychiatric Symptoms. Curr. Pharm. Des.
898 https://doi.org/10.2174/1381612821666150619092540
899 Frodl, T., Scheuerecker, J., Schoepf, V., Linn, J., Koutsouleris, N., Bokde, A.L.W.,
900 Hampel, H., Möller, H.J., Brückmann, H., Wiesmann, M., Meisenzahl, E., 2011.
901 Different effects of mirtazapine and venlafaxine on brain activation: An open
902 randomized controlled fMRI study. J. Clin. Psychiatry.

https://doi.org/10.4088/JCP.09m05393blu

Fu, C.H.Y., Costafreda, S.G., Sankar, A., Adams, T.M., Rasenick, M.M., Liu, P., Donati, R., Maglanoc, L.A., Horton, P., Marangell, L.B., 2015. Multimodal functional and structural neuroimaging investigation of major depressive disorder following treatment with duloxetine. BMC Psychiatry. https://doi.org/10.1186/s12888-015-0457-2

Fu, C.H.Y., Steiner, H., Costafreda, S.G., 2013. Predictive neural biomarkers of clinical response in depression: A meta-analysis of functional and structural neuroimaging studies of pharmacological and psychological therapies. Neurobiol. Dis. https://doi.org/10.1016/j.nbd.2012.05.008

Fu, C.H.Y., Williams, S.C.R., Brammer, M.J., Suckling, J., Kim, J., Cleare, A.J., Walsh, N.D., Mitterschiffthaler, M.T., Andrew, C.M., Pich, E.M., Bullmore, E.T., 2007. Neural responses to happy facial expressions in major depression following antidepressant treatment. Am. J. Psychiatry. https://doi.org/10.1176/ajp.2007.164.4.599

Fu, C.H.Y., Williams, S.C.R., Cleare, A.J., Brammer, M.J., Walsh, N.D., Kim, J., Andrew, C.M., Pich, E.M., Williams, P.M., Reed, L.J., Mitterschiffthaler, M.T., Suckling, J., Bullmore, E.T., 2004. Attenuation of the neural response to sad faces in major depression by antidepressant treatment: A prospective, event-related functional magnetic resonance imaging study. Arch. Gen. Psychiatry. https://doi.org/10.1001/archpsyc.61.9.877

Fu, C.H.Y., Williams, S.C.R., Cleare, A.J., Scott, J., Mitterschiffthaler, M.T., Walsh, N.D., Donaldson, C., Suckling, J., Andrew, C., Steiner, H., Murray, R.M., 2008. Neural Responses to Sad Facial Expressions in Major Depression Following Cognitive Behavioral Therapy. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2008.04.033

Furey, M.L., Drevets, W.C., Hoffman, E.M., Frankel, E., Speer, A.M., Zarate, C.A., 2013. Potential of pretreatment neural activity in the visual cortex during emotional processing to predict treatment response to scopolamine in major depressive disorder. JAMA Psychiatry. https://doi.org/10.1001/2013.jamapsychiatry.60

Furey, M.L., Drevets, W.C., Szczepanik, J., Khanna, A., Nugent, A., Zarate, C.A., 2015. Pretreatment differences in BOLD response to emotional faces correlate with Antidepress ant response to scopolamine. Int. J. Neuropsychopharmacol.

https://doi.org/10.1093/ijnp/pyv028

Galbaud Du Fort, G., Newman, S.C., Boothroyd, L.J., Bland, R.C., 1999. Treatment seeking for depression: Role of depressive symptoms and comorbid psychiatric diagnoses. J. Affect. Disord. https://doi.org/10.1016/S0165-0327(98)00052-4

Godlewska, B.R., Browning, M., Norbury, R., Cowen, P.J., Harmer, C.J., 2016. Early changes in emotional processing as a marker of clinical response to SSRI treatment in depression. Transl. Psychiatry. https://doi.org/10.1038/tp.2016.130

Godlewska, B.R., Browning, M., Norbury, R., Igoumenou, A., Cowen, P.J., Harmer, C.J., 2018. Predicting Treatment Response in Depression: The Role of Anterior Cingulate Cortex. Int. J. Neuropsychopharmacol. https://doi.org/10.1093/ijnp/pyy069

Godlewska, B.R., Norbury, R., Selvaraj, S., Cowen, P.J., Harmer, C.J., 2012. Short-term SSRI treatment normalises amygdala hyperactivity in depressed patients. Psychol. Med. https://doi.org/10.1017/S0033291712000591

Guo, C.C., Kurth, F., Zhou, J., Mayer, E.A., Eickhoff, S.B., Kramer, J.H., Seeley, W.W., 2012. One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. Neuroimage. https://doi.org/10.1016/j.neuroimage.2012.03.027

Gyurak, A., Patenaude, B., Korgaonkar, M.S., Grieve, S.M., Williams, L.M., Etkin, A., 2016. Frontoparietal activation during response inhibition predicts remission to antidepressants in patients with major depression. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2015.02.037

Hamilton, J.P., Glover, G.H., Bagarinao, E., Chang, C., Mackey, S., Sacchet, M.D., Gotlib, I.H., 2016. Effects of salience-network-node neurofeedback training on affective biases in major depressive disorder. Psychiatry Res. - Neuroimaging. https://doi.org/10.1016/j.pscychresns.2016.01.016

Handwerker, D.A., Gonzalez-Castillo, J., D'Esposito, M., Bandettini, P.A., 2012. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. Neuroimage. https://doi.org/10.1016/j.neuroimage.2012.02.015

Hansen, N.S., Siegle, G., 2015. Paving the road to the neurocognitive clinic of tomorrow: Appealing to standards, in: From Symptom to Synapse: A Neurocognitive Perspective on Clinical Psychology. https://doi.org/10.4324/9780203507131

Harrington, G.S., Tomaszewski Farias, S., Buonocore, M.H., Yonelinas, A.P., 2006. The intersubject and intrasubject reproducibility of FMRI activation during three

encoding tasks: Implications for clinical applications. Neuroradiology. https://doi.org/10.1007/s00234-006-0083-2

Heller, A.S., Johnstone, T., Light, S.N., Peterson, M.J., Kolden, G.G., Kalin, N.H., Davidson, R.J., 2013. Relationships between changes in sustained fronto-striatal connectivity and positive affect in major depression resulting from antidepressant treatment. Am. J. Psychiatry. https://doi.org/10.1176/appi.ajp.2012.12010014

Hrybouski, S., Aghamohammadi-Sereshki, A., Madan, C.R., Shafer, A.T., Baron, C.A., Seres, P., Beaulieu, C., Olsen, F., Malykhin, N. V., 2016. Amygdala subnuclei response and connectivity during emotional processing. Neuroimage. https://doi.org/10.1016/j.neuroimage.2016.02.056

Hsiao, C.K., Chen, P.C., Kao, W.H., 2011. Bayesian random effects for interrater and test-retest reliability with nested clinical observations. J. Clin. Epidemiol. https://doi.org/10.1016/j.jclinepi.2010.10.015

Hunter, A.M., Korb, A.S., Cook, I.A., Leuchter, A.F., 2013. Rostral anterior cingulate activity in major depressive disorder: State or trait marker of responsiveness to medication? J. Neuropsychiatry Clin. Neurosci. https://doi.org/10.1176/appi.neuropsych.11110330

Janak, P.H., Tye, K.M., 2015. From circuits to behaviour in the amygdala. Nature. https://doi.org/10.1038/nature14188

Johns, M.W., 1991. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. Sleep. https://doi.org/10.1093/sleep/14.6.540

Keedwell, P.A., Drapier, D., Surguladze, S., Giampietro, V., Brammer, M., Phillips, M., 2010. Subgenual cingulate and visual cortex responses to sad faces predict clinical outcome during antidepressant treatment for depression. J. Affect. Disord. https://doi.org/10.1016/j.jad.2009.04.031

Koenigs, M., Grafman, J., 2009. The functional neuroanatomy of depression: Distinct roles for ventromedial and dorsolateral prefrontal cortex. Behav. Brain Res. https://doi.org/10.1016/j.bbr.2009.03.004

Krüger, G., Glover, G.H., 2001. Physiological noise in oxygenation-sensitive magnetic resonance imaging. Magn. Reson. Med. https://doi.org/10.1002/mrm.1240

Laenen, A., Vangeneugden, T., Geys, H., Molenberghs, G., 2006. Generalized reliability estimation using repeated measurements. Br. J. Math. Stat. Psychol. https://doi.org/10.1348/000711005X66068

Lahey, M.A., Downey, R.G., Saal, F.E., 1983. Intraclass correlations: There's more there than meets the eye. Psychol. Bull. https://doi.org/10.1037/0033-2909.93.3.586

Langenecker, S.A., Kennedy, S.E., Guidotti, L.M., Briceno, E.M., Own, L.S., Hooven, T., Young, E.A., Akil, H., Noll, D.C., Zubieta, J.K., 2007. Frontal and Limbic Activation During Inhibitory Control Predicts Treatment Response in Major Depressive Disorder. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2007.02.019

Larson, C.L., Schaefer, H.S., Siegle, G.J., Jackson, C.A.B., Anderle, M.J., Davidson, R.J., 2006. Fear Is Fast in Phobic Individuals: Amygdala Activation in Response to Fear-Relevant Stimuli. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2006.03.079

Lebow, M.A., Chen, A., 2016. Overshadowed by the amygdala: The bed nucleus of the stria terminalis emerges as key to psychiatric disorders. Mol. Psychiatry. https://doi.org/10.1038/mp.2016.1

Lecrux, C., Bourourou, M., Hamel, E., 2019. How reliable is cerebral blood flow to map changes in neuronal activity? Auton. Neurosci. Basic Clin. https://doi.org/10.1016/j.autneu.2019.01.005

LeDoux, J., 2012. Rethinking the Emotional Brain. Neuron. https://doi.org/10.1016/j.neuron.2012.02.004

Lemogne, C., Mayberg, H., Bergouignan, L., Volle, E., Delaveau, P., Lehéricy, S., Allilaire, J.F., Fossati, P., 2010. Self-referential processing and the prefrontal cortex over the course of depression: A pilot study. J. Affect. Disord. https://doi.org/10.1016/j.jad.2009.11.003

Light, S.N., Heller, A.S., Johnstone, T., Kolden, G.G., Peterson, M.J., Kalin, N.H., Davidson, R.J., 2011. Reduced right ventrolateral prefrontal cortex activity while inhibiting positive affect is associated with improvement in hedonic capacity after 8 weeks of antidepressant treatment in major depressive disorder. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2011.06.031

Lin, A.L., Monica Way, H.Y., 2014. Functional Magnetic Resonance Imaging, in: Pathobiology of Human Disease: A Dynamic Encyclopedia of Disease Mechanisms. https://doi.org/10.1016/B978-0-12-386456-7.07610-3

Linden, D.E.J., 2014. Neurofeedback and networks of depression. Dialogues Clin. Neurosci.

1039 Linden, D.E.J., Habes, I., Johnston, S.J., Linden, S., Tatineni, R., Subramanian, L.,
1040 Sorger, B., Healy, D., Goebel, R., 2012. Real-time self-regulation of emotion
1041 networks in patients with depression. PLoS One.
1042 https://doi.org/10.1371/journal.pone.0038115
1043 Lindquist, M.A., Meng Loh, J., Atlas, L.Y., Wager, T.D., 2009. Modeling the
1044 hemodynamic response function in fMRI: efficiency, bias and mis-modeling.
1045 Neuroimage. https://doi.org/10.1016/j.neuroimage.2008.10.065
1046 Lois, G., Kirsch, P., Sandner, M., Plichta, M.M., Wessa, M., 2018. Experimental and
1047 methodological factors affecting test-retest reliability of amygdala BOLD
1048 responses. Psychophysiology. https://doi.org/10.1111/psyp.13220
1049 López-Solà, M., Pujol, J., Hernández-Ribas, R., Harrison, B.J., Contreras-Rodríguez,
1050 O., Soriano-Mas, C., Deus, J., Ortiz, H., Menchón, J.M., Vallejo, J., Cardoner,
1051 N., 2010. Effects of duloxetine treatment on brain response to painful stimulation
1052 in major depressive disorder. Neuropsychopharmacology.
1053 https://doi.org/10.1038/npp.2010.108
1054 MacDuffie, K.E., MacInnes, J., Dickerson, K.C., Eddington, K.M., Strauman, T.J.,
1055 Adcock, R.A., 2018. Single session real-time fMRI neurofeedback has a lasting
1056 impact on cognitive behavioral therapy strategies. NeuroImage Clin.
1057 https://doi.org/10.1016/j.nicl.2018.06.009
1058 Mandell, D., Siegle, G.J., Shutt, L., Feldmiller, J., Thase, M.E., 2014. Neural
1059 substrates of trait ruminations in depression. J. Abnorm. Psychol. 123, 35–48.
1060 https://doi.org/10.1037/a0035834
1061 Mehler, D.M.A., Sokunbi, M.O., Habes, I., Barawi, K., Subramanian, L., Range, M.,
1062 Evans, J., Hood, K., Lührs, M., Keedwell, P., Goebel, R., Linden, D.E.J., 2018.
1063 Targeting the affective brain—a randomized controlled trial of real-time fMRI
1064 neurofeedback in patients with depression. Neuropsychopharmacology.
1065 https://doi.org/10.1038/s41386-018-0126-5
1066 Michely, J., Rigoli, F., Rutledge, R.B., Hauser, T.U., Dolan, R.J., 2020. Distinct
1067 Processing of Aversive Experience in Amygdala Subregions. Biol. Psychiatry
1068 Cogn. Neurosci. Neuroimaging 5, 291–300.
1069 https://doi.org/10.1016/j.bpsc.2019.07.008
1070 Miki, A., Raz, J., Van Erp, T.G.M., Liu, C.S.J., Haselgrove, J.C., Liu, G.T., 2000.
1071 Reproducibility of visual activation in functional MR imaging and effects of
1072 postprocessing. Am. J. Neuroradiol.

1073 Miller, J.M., Schneck, N., Siegle, G.J., Chen, Y., Ogden, R.T., Kikuchi, T., Oquendo,
1074 M.A., Mann, J.J., Parsey, R. V., 2013. FMRI response to negative words and
1075 SSRI treatment outcome in major depressive disorder: A preliminary study.
1076 Psychiatry Res. - Neuroimaging.
1077 https://doi.org/10.1016/j.pscychresns.2013.08.001
1078 Nichols, T.E., Das, S., Eickhoff, S.B., Evans, A.C., Glatard, T., Hanke, M.,
1079 Kriegeskorte, N., Milham, M.P., Poldrack, R.A., Poline, J.B., Proal, E., Thirion,
1080 B., Van Essen, D.C., White, T., Yeo, B.T.T., 2017. Best practices in data
1081 analysis and sharing in neuroimaging using MRI. Nat. Neurosci.
1082 https://doi.org/10.1038/nn.4500
1083 Noble, S., Scheinost, D., Constable, R.T., 2019. A decade of test-retest reliability of
1084 functional connectivity: A systematic review and meta-analysis. Neuroimage.
1085 https://doi.org/10.1016/j.neuroimage.2019.116157
1086 Nolen-Hoeksema, S., Morrow, J., Fredrickson, B.L., 1993. Response styles and the
1087 duration of episodes of depressed mood. J. Abnorm. Psychol. 102, 20–28.
1088 https://doi.org/10.1037/0021-843X.102.1.20
1089 Nord, C.L., Gray, A., Charpentier, C.J., Robinson, O.J., Roiser, J.P., 2017.
1090 Unreliability of putative fMRI biomarkers during emotional face processing.
1091 Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.05.024
1092 Oakes, T.R., Johnstone, T., Ores Walsh, K.S., Greischar, L.L., Alexander, A.L., Fox,
1093 A.S., Davidson, R.J., 2005. Comparison of fMRI motion correction software
1094 tools. Neuroimage. https://doi.org/10.1016/j.neuroimage.2005.05.058
1095 Opmeer, E.M., Kortekaas, R., Van Tol, M.J., Renken, R.J., Demenescu, L.R.,
1096 Woudstra, S., Ter Horst, G.J., Van Buchem, M.A., Van Der Wee, N.J.A.,
1097 Veltman, D.J., Aleman, A., 2016. Changes in regional brain activation related to
1098 depressive state: A 2-year longitudinal functional MRI study. Depress. Anxiety.
1099 https://doi.org/10.1002/da.22425
1100 Paek, E.J., Murray, L.L., Newman, S.D., Kim, D.J., 2019. Test-retest reliability in an
1101 fMRI study of naming in dementia. Brain Lang.
1102 https://doi.org/10.1016/j.bandl.2019.02.002
1103 Palmer, C.E., Langbehn, D., Tabrizi, S.J., Papoutsi, M., 2018. Test-retest reliability of
1104 measures commonly used to measure striatal dysfunction across multiple testing
1105 sessions: A longitudinal study. Front. Psychol.
1106 https://doi.org/10.3389/fpsyg.2017.02363

Palomero-Gallagher, N., Hoffstaedter, F., Mohlberg, H., Eickhoff, S.B., Amunts, K., Zilles, K., 2019. Human Pregenual Anterior Cingulate Cortex: Structural, Functional, and Connectional Heterogeneity. Cereb. Cortex. https://doi.org/10.1093/cercor/bhy124

Peng, S.-L., Chen, C.-M., Huang, C.-Y., Shih, C.-T., Huang, C.-W., Chiu, S.-C., Shen, W.-C., 2019. Effects of Hemodynamic Response Function Selection on Rat fMRI Statistical Analyses. Front. Neurosci. https://doi.org/10.3389/fnins.2019.00400

Phillips, M.L., Swartz, H.A., 2014. A Critical Appraisal of Neuroimaging Studies of Bipolar Disorder: Toward a New Conceptualization of Underlying Neural Circuitry and a Road Map for Future Research. Am. J. Psychiatry 171, 829–843. https://doi.org/10.1176/appi.ajp.2014.13081008

Pickford, R.W., Guilford, J.P., 2007. Psychometric Methods. Br. J. Sociol. https://doi.org/10.2307/586971

Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A.B.M., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. Neuroimage. https://doi.org/10.1016/j.neuroimage.2012.01.129

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: Towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. https://doi.org/10.1038/nrn.2016.167

Portney, L., Watkins, M., 2009. Foundations of Clinical Research: Applications to Practice, Upper Saddle River, NJ: Pearson/Prentice Hall.

Ritchey, M., Dolcos, F., Eddington, K.M., Strauman, T.J., Cabeza, R., 2011. Neural correlates of emotional processing in depression: Changes with cognitive behavioral therapy and predictors of treatment response. J. Psychiatr. Res. https://doi.org/10.1016/j.jpsychires.2010.09.007

Rizvi, S.J., Salomons, T. V., Konarski, J.Z., Downar, J., Giacobbe, P., McIntyre, R.S., Kennedy, S.H., 2013. Neural response to emotional stimuli associated with successful antidepressant treatment and behavioral activation. J. Affect. Disord. https://doi.org/10.1016/j.jad.2013.06.050

Robertson, B., Wang, L., Diaz, M.T., Aiello, M., Gersing, K., Beyer, J., Mukundan, S.,

McCarthy, G., Doraiswamy, P.M., 2007. Effect of bupropion extended release on negative emotion processing in major depressive disorder: A pilot functional magnetic resonance imaging study. J. Clin. Psychiatry. https://doi.org/10.4088/JCP.v68n0212

Rosenblau, G., Sterzer, P., Stoy, M., Park, S., Friedel, E., Heinz, A., Pilhatsch, M., Bauer, M., Ströhle, A., 2012. Functional neuroanatomy of emotion processing in major depressive disorder is altered after successful antidepressant therapy. J. Psychopharmacol. https://doi.org/10.1177/0269881112450779

Roy, A.K., Shehzad, Z., Margulies, D.S., Kelly, A.M.C., Uddin, L.Q., Gotimer, K., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2009. Functional connectivity of the human amygdala using resting state fMRI. Neuroimage. https://doi.org/10.1016/j.neuroimage.2008.11.030

Roy, M., Harvey, P.O., Berlim, M.T., Mamdani, F., Beaulieu, M.M., Turecki, G., Lepage, M., 2010. Medial prefrontal cortex activity during memory encoding of pictures and its relation to symptomatic improvement after citalopram treatment in patients with major depression. J. Psychiatry Neurosci. https://doi.org/10.1503/jpn.090010

Rubin-Falcone, H., Weber, J., Kishon, R., Ochsner, K., Delaparte, L., Doré, B., Zanderigo, F., Oquendo, M.A., Mann, J.J., Miller, J.M., 2018. Longitudinal effects of cognitive behavioral therapy for depression on the neural correlates of emotion regulation. Psychiatry Res. - Neuroimaging. https://doi.org/10.1016/j.pscychresns.2017.11.002

Ruhé, H.G., Booij, J., Veltman, D.J., Michel, M.C., Schene, A.H., 2012. Successful pharmacologic treatment of major depressive disorder attenuates amygdala activation to negative facial expressions: A functional magnetic resonance imaging study. J. Clin. Psychiatry. https://doi.org/10.4088/JCP.10m06584

Ruiz, S., Buyukturkoglu, K., Rana, M., Birbaumer, N., Sitaram, R., 2014. Real-time fMRI brain computer interfaces: Self-regulation of single brain regions to networks. Biol. Psychol. https://doi.org/10.1016/j.biopsycho.2013.04.010

Samson, A.C., Meisenzahl, E., Scheuerecker, J., Rose, E., Schoepf, V., Wiesmann, M., Frodl, T., 2011. Brain activation predicts treatment improvement in patients with major depressive disorder. J. Psychiatr. Res. https://doi.org/10.1016/j.jpsychires.2011.03.009

Sankar, A., Adams, T.M., Costafreda, S.G., Marangell, L.B., Fu, C.H.Y., 2017.

1175 Effects of antidepressant therapy on neural components of verbal working
1176 memory in depression. J. Psychopharmacol.
1177 https://doi.org/10.1177/0269881117724594
1178 Schaefer, H.S., Putnam, K.M., Benca, R.M., Davidson, R.J., 2006. Event-Related
1179 Functional Magnetic Resonance Imaging Measures of Neural Activity to Positive
1180 Social Stimuli in Pre- and Post-Treatment Depression. Biol. Psychiatry.
1181 https://doi.org/10.1016/j.biopsych.2006.03.024
1182 Shan, Z.Y., Wright, M.J., Thompson, P.M., McMahon, K.L., Blokland, G.G.A.M., De
1183 Zubicaray, G.I., Martin, N.G., Vinkhuyzen, A.A.E., Reutens, D.C., 2014.
1184 Modeling of the hemodynamic responses in block design fMRI studies. J. Cereb.
1185 Blood Flow Metab. https://doi.org/10.1038/jcbfm.2013.200
1186 Sheatsley, P.B., 1983. Questionnaire Construction and Item Writing, in: Handbook of
1187 Survey Research. https://doi.org/10.1016/b978-0-12-598226-9.50012-4
1188 Sheline, Y.I., Barch, D.M., Donnelly, J.M., Ollinger, J.M., Snyder, A.Z., Mintun, M.A.,
1189 2001. Increased amygdala response to masked emotional faces in depressed
1190 subjects resolves with antidepressant treatment: An fMRI study. Biol. Psychiatry.
1191 https://doi.org/10.1016/S0006-3223(01)01263-X
1192 Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater
1193 reliability. Psychol. Bull. https://doi.org/10.1037/0033-2909.86.2.420
1194 Siegle, G.J., Carter, C.S., Thase, M.E., 2006. Use of fMRI to predict recovery from
1195 unipolar depression with cognitive behavior therapy. Am. J. Psychiatry.
1196 https://doi.org/10.1176/ajp.2006.163.4.735
1197 Siegle, G.J., Steinhauer, S.R., Thase, M.E., Stenger, V.A., Carter, C.S., 2002. Can't
1198 shake that feeling: Event-related fMRI assessment of sustained amygdala
1199 activity in response to emotional information in depressed individuals. Biol.
1200 Psychiatry. https://doi.org/10.4324/9780203706978
1201 Siegle, G.J., Thompson, W.K., Collier, A., Berman, S.R., Feldmiller, J., Thase, M.E.,
1202 Friedman, E.S., 2012. Toward Clinically Useful Neuroimaging in Depression
1203 Treatment. Arch. Gen. Psychiatry.
1204 https://doi.org/10.1001/archgenpsychiatry.2012.65
1205 Sladky, R., Höflich, A., Atanelov, J., Kraus, C., Baldinger, P., Moser, E.,
1206 Lanzenberger, R., Windischberger, C., 2012. Increased Neural Habituation in
1207 the Amygdala and Orbitofrontal Cortex in Social Anxiety Disorder Revealed by
1208 fMRI. PLoS One. https://doi.org/10.1371/journal.pone.0050050

1209 Spielberger, C., 1983. Manual for the State-Trait Anxiety Inventory (STAI). Consult.
1210 Psychol. Press 4–26. https://doi.org/10.1007/978-0-387-78665-0_6709
1211 Spies, M., Kraus, C., Geissberger, N., Auer, B., Klöbl, M., Tik, M., Störkat, I.L., Hahn,
1212 A., Woletz, M., Pfabigan, D.M., Kasper, S., Lamm, C., Windischberger, C.,
1213 Lanzenberger, R., 2017. Default mode network deactivation during emotion
1214 processing predicts early antidepressant response. Transl. Psychiatry.
1215 https://doi.org/10.1038/tp.2016.265
1216 Stoy, M., Schlagenhauf, F., Sterzer, P., Bermpohl, F., Hägele, C., Suchotzki, K.,
1217 Schmack, K., Wrase, J., Ricken, R., Knutson, B., Adli, M., Bauer, M., Heinz, A.,
1218 Ströhle, A., 2012. Hyporeactivity of ventral striatum towards incentive stimuli in
1219 unmedicated depressed patients normalizes after treatment with escitalopram. J.
1220 Psychopharmacol. https://doi.org/10.1177/0269881111416686
1221 Straub, J., Plener, P.L., Sproeber, N., Sprenger, L., Koelch, M.G., Groen, G., Abler,
1222 B., 2015. Neural correlates of successful psychotherapy of depression in
1223 adolescents. J. Affect. Disord. https://doi.org/10.1016/j.jad.2015.05.020
1224 Strege, M. V, Siegle, G.J., Young, K., 2020. Cingulate prediction of response to
1225 antidepressant and cognitive behavioral therapies for depression: Theory, meta-
1226 analysis, and empirical application. bioRxiv 2020.12.02.407841.
1227 https://doi.org/10.1101/2020.12.02.407841
1228 Szczepanik, J., Nugent, A.C., Drevets, W.C., Khanna, A., Zarate, C.A., Furey, M.L.,
1229 2016. Amygdala response to explicit sad face stimuli at baseline predicts
1230 antidepressant treatment response to scopolamine in major depressive disorder.
1231 Psychiatry Res. - Neuroimaging.
1232 https://doi.org/10.1016/j.pscychresns.2016.06.005
1233 Tao, R., Calley, C.S., Hart, J., Mayes, T.L., Nakonezny, P.A., Lu, H., Kennard, B.D.,
1234 Tamminga, C.A., Emslie, G.J., 2012. Brain activity in adolescent major
1235 depressive disorder before and after fluoxetine treatment. Am. J. Psychiatry.
1236 https://doi.org/10.1176/appi.ajp.2011.11040615
1237 Taylor, S.F., Ho, S.S., Abagis, T., Angstadt, M., Maixner, D.F., Welsh, R.C.,
1238 Hernandez-Garcia, L., 2018. Changes in brain connectivity during a sham-
1239 controlled, transcranial magnetic stimulation trial for depression. J. Affect.
1240 Disord. https://doi.org/10.1016/j.jad.2018.02.019
1241 Thibault, R.T., MacPherson, A., Lifshitz, M., Roth, R.R., Raz, A., 2018.
1242 Neurofeedback with fMRI: A critical systematic review. Neuroimage.

https://doi.org/10.1016/j.neuroimage.2017.12.071

Toki, S., Okamoto, Y., Onoda, K., Matsumoto, T., Yoshimura, S., Kunisato, Y., Okada, G., Shishida, K., Kobayakawa, M., Fukumoto, T., Machino, A., Inagaki, M., Yamawaki, S., 2014. Hippocampal activation during associative encoding of word pairs and its relation to symptomatic improvement in depression: A functional and volumetric MRI study. J. Affect. Disord. https://doi.org/10.1016/j.jad.2013.07.021

Victor, T.A., Furey, M.L., Fromm, S.J., Öhman, A., Drevets, W.C., 2013. Changes in the neural correlates of implicit emotional face processing during antidepressant treatment in major depressive disorder, in: International Journal of Neuropsychopharmacology. https://doi.org/10.1017/S146114571300062X

Victor, T.A., Furey, M.L., Fromm, S.J., Öhman, A., Drevets, W.C., 2010. Relationship between amygdala responses to masked faces and mood state and treatment in major depressive disorder. Arch. Gen. Psychiatry. https://doi.org/10.1001/archgenpsychiatry.2010.144

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition (a.k.a. Voodoo Correlations in Social Neuroscience). Perspect. Psychol. Sci.

Wagner, G., Koch, K., Schachtzabel, C., Sobanski, T., Reichenbach, J.R., Sauer, H., Schlösser, R.G.M., 2010. Differential effects of serotonergic and noradrenergic antidepressants on brain activity during a cognitive control task and neurofunctional prediction of treatment outcome in patients with depression. J. Psychiatry Neurosci. https://doi.org/10.1503/jpn.090081

Walsh, N.D., Williams, S.C.R., Brammer, M.J., Bullmore, E.T., Kim, J., Suckling, J., Mitterschiffthaler, M.T., Cleare, A.J., Pich, E.M., Mehta, M.A., Fu, C.H.Y., 2007. A Longitudinal Functional Magnetic Resonance Imaging Study of Verbal Working Memory in Depression After Antidepressant Therapy. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2006.12.022

Wang, Y., Xu, C., Cao, X., Gao, Q., Li, J., Liu, Z., Sun, N., Ren, Y., Zhang, K., 2012. Effects of an antidepressant on neural correlates of emotional processing in patients with major depression. Neurosci. Lett. https://doi.org/10.1016/j.neulet.2012.08.034

Wessa, M., Lois, G., 2015. Brain Functional Effects of Psychopharmacological Treatment in Major Depression: a Focus on Neural Circuitry of Affective

Processing. Curr. Neuropharmacol. https://doi.org/10.2174/1570159x13666150416224801

Williams, L.M., Korgaonkar, M.S., Song, Y.C., Paton, R., Eagles, S., Goldstein-Piekarski, A., Grieve, S.M., Harris, A.W.F., Usherwood, T., Etkin, A., 2015. Amygdala Reactivity to Emotional Faces in the Prediction of General and Medication-Specific Responses to Antidepressant Treatment in the Randomized iSPOT-D Trial. Neuropsychopharmacology. https://doi.org/10.1038/npp.2015.89

Yoshimura, S., Okamoto, Y., Onoda, K., Matsunaga, M., Okada, G., Kunisato, Y., Yoshino, A., Ueda, K., Suzuki, S. ichi, Yamawaki, S., 2014. Cognitive behavioral therapy for depression changes medial prefrontal and ventral anterior cingulate cortex activity associated with self-referential processing. Soc. Cogn. Affect. Neurosci. https://doi.org/10.1093/scan/nst009

Young, K.D., Misaki, M., Harmer, C.J., Victor, T., Zotev, V., Phillips, R., Siegle, G.J., Drevets, W.C., Bodurka, J., 2017a. Real-Time fMRI Amygdala Neurofeedback Changes Positive Information Processing in Major Depressive Disorder. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2017.03.013

Young, K.D., Siegle, G.J., Misaki, M., Zotev, V., Phillips, R., Drevets, W.C., Bodurka, J., 2018. Altered task-based and resting-state amygdala functional connectivity following real-time fMRI amygdala neurofeedback training in major depressive disorder. NeuroImage Clin. https://doi.org/10.1016/j.nicl.2017.12.004

Young, K.D., Siegle, G.J., Zotev, V., Phillips, R., Misaki, M., Yuan, H., Drevets, W.C., Bodurka, J., 2017b. Randomized clinical trial of real-time fMRI amygdala neurofeedbackfor major depressive disorder: Effectson symptoms and autobiographical memory recall, in: American Journal of Psychiatry. https://doi.org/10.1176/appi.ajp.2017.16060637

Young, K.D., Zotev, V., Phillips, R., Misaki, M., Yuan, H., Drevets, W.C., Bodurka, J., 2014. Real-time fMRI neurofeedback training of amygdala activity in patients with major depressive disorder. PLoS One. https://doi.org/10.1371/journal.pone.0088785

Yuan, H., Young, K.D., Phillips, R., Zotev, V., Misaki, M., Bodurka, J., 2014. Resting-State Functional Connectivity Modulation and Sustained Changes After Real-Time Functional Magnetic Resonance Imaging Neurofeedback Training in Depression. Brain Connect. https://doi.org/10.1089/brain.2014.0262

Zandbelt, B.B., Gladwin, T.E., Raemaekers, M., van Buuren, M., Neggers, S.F.,

Kahn, R.S., Ramsey, N.F., Vink, M., 2008. Within-subject variation in BOLD-fMRI signal changes across repeated measurements: Quantification and implications for sample size. Neuroimage. https://doi.org/10.1016/j.neuroimage.2008.04.183

Zhilkin, P., Alexander, M.E., 2004. Affine registration: A comparison of several programs. Magn. Reson. Imaging. https://doi.org/10.1016/j.mri.2003.05.004

Zotev, V., Phillips, R., Yuan, H., Misaki, M., Bodurka, J., 2014. Self-regulation of human brain activity using simultaneous real-time fMRI and EEG neurofeedback. Neuroimage. https://doi.org/10.1016/j.neuroimage.2013.04.126

Zotev, V., Yuan, H., Misaki, M., Phillips, R., Young, K.D., Feldner, M.T., Bodurka, J., 2016. Correlation between amygdala BOLD activity and frontal EEG asymmetry during real-time fMRI neurofeedback training in patients with depression. NeuroImage Clin. https://doi.org/10.1016/j.nicl.2016.02.003