Towards clinically useful neuroimaging in depression treatment: Is subgenual cingulate activity robustly prognostic for depression outcome in Cognitive Therapy across studies, scanners, and patient characteristics?

Greg J. Siegle, Ph.D., Wesley Thompson, Ph.D., Amanda Collier, B.A.,Susan Berman, R.N., Joshua Feldmiller, B.A., Michael E. Thase, M.D., Edward S. Friedman, M.D.

Author Material

Contents

Author Material-I. Description of Cohorts and Cognitive Therapy in Each Cohort

***I-A. Cohort Definition and allocation*.** The data in this study were analyzed as two cohorts separated by both a change in scanner and clinical trial status. The formal delimiter was the scanner on which they were run given strong concerns in the field regarding combining data across scanners and the confounding of clinical-trial and scanner within this study. Specifically, in 2006 the University of Pittsburgh decommissioned the GE 3T magnet on which the first cohort was run, and we installed a new Siemmen's Trio magnet. Within 3 months of that switch, recruitment for the first associated clinical trial (Thase, PI) ended. Six months later, recruitment for the second associated clinical trial (Siegle, PI) began. Thus, three patients from the Thase et al trial were run on the new scanner. Given the high level of heterogeneity inherent in the design of the Siegle et al effectiveness trial, including these three patients in the otherwise heterogeneous mix of this trial seemed a reasonable definition. Sensitivity analyses excluding these patients did not suggest that they were qualitatively different from the rest of the sample.

I-B. Cognitive Therapy. The employed Cognitive Therapy followed Beck's (1) guidelines. The primary focus was on identifying thought/feeling relationships, followed by learning skills for challenging negative thoughts and adopting more adaptive thoughts. Cognitive and behavioral exercises were frequently prescribed for homework.

All patients received the flyer "Questions and Answers about Cognitive Therapy" (2). Patients were frequently, but not always, directed to purchase and use the workbook *Mind Over Mood (3).*

I-C. Supervision. In Cohort 1, therapists received weekly supervision accompanied by a Cognitive Therapy Rating Scale (4) evaluation. Scores of 40 and above were considered adequate therapy delivery. Supervision was oriented towards improving adherence to the Cognitive Therapy model by improving rating scale scores. Therapists who scored below this range were not retained. In Cohort 2 the rating scale was used descriptively with a target of 40, but therapists were not generally shown their scores to preserve the community practitioner feel of the study. Supervision was still oriented towards improving adherence.

Author Material-II. Clinical Data Preparation

II-A. Rationale for use of BDI as primary outcome measure but inclusion of HRSD. We chose the BDI as our primary measure because the initial goal of this work was to replicate our previous observation in (5) which involved consideration of residual severity measured using the BDI. That said, as the HRSD is the "industry standard", we have reported on this measure as well. The specific reason for choosing the BDI as an outcome measure in our previous study is that we were specifically examining response to Cognitive Therapy which targets cognitive symptoms of depression. Cognitive symptoms are more strongly represented on the BDI than on the HRSD.

II-B. Reconstruction of missing BDI-II scores from BDI-I data. For pre-treatment scores, the BDI-I and II strongly corresponded, R^2=0.62, F(1,20)=32.78, p<0.0005, avg pts off = 3.86, with the match for computed as BDI-II=13.59+0.73BDI-I. For post-treatment, scores also corresponded strongly, R^2=0.88, F(1,16)=113.03, p=0.000, avg pts off = 2.33 with the match computed as BDI-II=0.75+1.19BDI-I.

II-C. Rationale for BDI response threshold. While there is no agreed-on threshold for remission based on the BDI (6), criteria have ranged from <9 (7) to <12 (8); we opted for a criterion more heavily weighted towards specificity as the probable use of this measure is likely to be in deciding to opt out of, rather than into therapy.

II-D. Imputation of final HRSD responses. As weekly HRSD scores were obtained, it was possible to estimate final HRSD scores consistent with participants' response trajectory independent of minor reporting changes in the last weeks not associated with true treatment response. Towards that end we used an imputation procedure for final HRSD scores using a functional linear model. Each person's time-varying depression scores were modeled as a linear combination of B-splines with random (multivariate normal) coefficients. The level of smoothness was data-determined by estimating the variance of the random coefficients. Final values were imputed by assessing the expected value from the B-spline regression at 16 weeks. Imputed scores were strongly correlated with the minimum Hamilton score for each participant, r=.79, p<.005, M(points different)=2.71, and a.so with the last valid Hamilton score, r=.89, p<.005, M(points different)=2.38.

Author Material-III. Procedures

III-A. Imaging task battery. The counterbalanced tasks, which will be reported on in separate publications, included a personal relevance rating task (described here), digit sorting task, emotional face viewing, and automatic thought rating. A second set of other tasks not described here was administered after the first set.

III-B. Acquisition of personally relevant words. Ten normed positive, 10 negative, and 10 neutral words (Cohort 1: trait adjectives, Cohort 2: trait adjectives for 19 patients 15 controls, or mixed nouns/adjectives as in previous studies for 5 controls, 13 patients) balanced for arousal, normed affect, word frequency, and word length were chosen using a computer program (9) that drew words from the ANEW (10) master list. Before the experiment, participants also generated "10 personally relevant negative words that best represent what you think about when you are upset, down, or depressed," "10 personally relevant positive words that best represent what you think about when you are happy or in a good mood," and "10 personally relevant neutral (i.e., not positive or negative) words that best represent what you think about when you are neither very happy nor very upset, down, or depressed."

Author Material-IV. fMRI data preparation

IV-A. fMRI data Preparation. fMRI analyses were conducted via locally developed NeuroImaging Software (NIS) and AFNI (11). Following slice-time correction, motion correction (AFNI 3dVolReg) and linear detrending to eliminate scanner drift, voxelwise outliers were Winsorized as for RTs. fMRI data were converted to percent-change from the voxel's median within a run, temporally smoothed (five-point middle-peaked filter), cross-registered to the Colin-27 Montreal Neurological Institute template (12) (AIR's 32-parameter non-linear warp; (13)), and spatially smoothed (6mm FWHM). Time-series variability was normalized across scanners. The same "reactivity" index used in (5) yielded peak and sustained responses to negative words as the mean of scans 4-7 of each negative-word trial minus the trial's first (pre-stimulus) scan. No model was applied given the potential for non-canonically shaped responses in the depressed group (e.g., 14, 15).

The inclusion of %-change as a pre-processing step was an explicit decision because we were concerned that for a possibly clinically applicable method, it was important that analyses apply exactly as specified to the preprocessed data, and thus, conversion to %-change was implemented as a voxelwise preprocessing step. Put another way, in general fMRI voxelwise analyses are conducted on the raw MR signal and then results are converted post-hoc to %-change which allows for the possibility that reported results don't match the analyses. Ours are guaranteed to match.

IV-B. fMRI Type I error control. To control type I error, voxelwise tests were thresholded at $p<.001$ and subjected to empirically derived contiguity thresholding via Monte-Carlo simulations accounting for the spatial autocorrelation of derived maps using AFNI's 3dFWHMx and 3dAlphaSim programs (11) (24 voxels yielding $p<.05$, corrected). Within the sgACC mask, one voxel at $p<.001$ controlled type 1 error at $p<.05$ via small-volume correction.

IV-C. Comparison of processing stream with our original 2006 publication (5). Differences in scanning between our previous study (5) and the current protocol reflected modernization's in our lab's processing stream not specific to this study. Extensive testing has shown this new processing stream to be slightly more robust than our previous stream. Changes included a faster motion correction algorithm (Afni's (11) 3dVolReg vs. AIR (16, 17) - extensive testing noted no differences) and modern standard preprocessing steps not used in (5) (slice time correction, multiplicative baseline correction to express %-change with respect to the time-series median before additive trial-based offsets were calculated, non-linear warping to a standard brain rather than linear transformation to a study brain).

IV-D. Method for accounting for multiple scanners: Cross-scanner variability normalization. As (5) used only one scanner there was no variance normalization across scanners in that study. Here, we equated the mean variability subject to outlier Windsorization as described in the text within-time series across participants, across scanners by scaling the timeseries for regions from Cohort 2 as Cohort2=Cohort2*(std(Cohort1)/std(Cohort2)). Thus, if the scanner used with Cohort 2 produced increased or decreased time-series variability compared to that used with Cohort 1 it would have been accounted for.

Of note, this scaling was used rather than scaling the mean response as we did not want to remove true additive cohort related effects which were analyzed. Moreover, scaling by the mean response is not only unprincipled (there is no reason why the mean of the time-series should scale without the variability) it is frequently misleading, e.g., if one scanner has a negative mean whereas the other has a positive mean the sign of the response can flip.

Importantly, for this study, scanner-related differences in the mean response magnitude (AUC for the response to word stimuli) were small both before and after the applied scaling correction (1.24). For the sgACC, before correction, $t(159)=-0.69$, $p=0.49$, M(SD)Scanner1=-.0038(.0619), M(SD)Scanner2=-.0096(.044), Mean Difference=-0.01(0.05), $d=-0.11$, and after correction, $t(159)=-0.87$, $p=0.39$, M(SD)Scanner1=-.0038(.0619), M(SD)Scanner2=-.0119(.055), Mean Difference =-0.01(0.06), $d=-0.14$. The variabilities were not exactly the same as outliers were Windsorized before computing the scale factor.

IV-E. Plan for analyzing scanner processing stream differences from (5)***.*** Our *a priori* approach to dealing with scanner-related differences was that if results differed from (5) we resolved to re-preprocess that data to understand differences, but if they were similar we resolved to keep the new stream as it better represents the current state of the art and supports robustness of the basic result to processing stream differences.

Author Material –V. A priori region selection

Goals in selecting *a priori* regions were to select regions that would 1) be consistent with the published literature and 2) reflect anatomical and functional homogeneity. When possible we used anatomically defined region-masks.

V-A. DLPFC. We used an empirically, rather than anatomically identified DLPFC region as the DLPFC encompasses large regions of potential functional heterogeneity, and because relevant subregions are reliably differentiated on exploratory analyses of tasks involving cognitive control and emotional information processing. We used a left mask from a region that previously differentiated depressed and control participants on digit sorting (15) – depressed participants displayed decreased activity in this region on the digit sorting task as well as on the Personal Relevance Rating task used in the current study. As the region was selected in one task and applied to another task, empirical region selection was not subject to common criticisms of capitalization on chance, yielding a likely robust region. To select a right-sided homologue that was also robust to capitalization on chance, we imposed the further constraint that the selected region must display parametric activation with task difficulty in both depressed and controls on the digit sorting task task via conjunction analysis (within-group ANOVAs thresholded at $p<0.05$ for each component map yielding $p<0.0025$ for the conjunction, thresholded empirically).

V-B. Amygdala. An anatomical left amygdala region was hand traced on the template via boundaries as in (5, 15) for consistency with prior work. We have previously shown that this anatomically defined region 1) differentiates healthy and depressed participants on the employed task (15) and contains voxels which predict response to CT on the employed task (5). We have established adequate intra- and inter-rater reliability (14) for this definition, with boundaries defined as: posterior: the alveus of the hippocampus, anterior: 2mm from the temporal horn of the lateral ventrical, superior: ventral horn of the sub-arachnoid space (SS), inferior: most dorsal finger of the white matter tract under the horn of the SS, lateral: 2mm from the surrounding white matter, mesial: 2mm from the SS).

This region definition differed minimally from a Talairach Atlas based version, with the primary differences being imposing a constraint of 1mm boundaries from the medial and anterior boundaries of the subarachnoid space, ensuring the non-inclusion of peri-amygdaloid cortex, as well as exclusion of extended amygdala regions such as the bed nucleus of the stria terminalus.

V-C. BA24. We defined an area of the rostral cingulate including BA24 using Afni's Talairach atlas. The rationale for selection of this anatomically defined region is that BA24 in the rostral cingulate consistently predicts response in medication studies using PET (18-20), fMRI (21-26), and EEG (27). That said, as the predictive areas have been derived empirically in all of these studies, and do not always overlap, we chose a boarder anatomically defined region which does encompass the majority of regions observed in previous studies and preserve cytoarchitectonic homogeneity.

Author Material-VI. Analysis of behavioral data

VI-A. Emotion Rating Effects. As shown in Table S1, Cohort 1 and Cohort 2 differed in their emotion ratings based on valence, Cohort x valence $F(2,79)=5.81, p=.004, \sim\eta^2=.0.13$. Cohort 2 reliably rated positive words as more positive $t(82)=-2.54$, $p=0.013$, $d=0.57$, and Cohort 2 consistently rated negative words as more negative $t(82)=3.27$, $p=0.002$, $d=0.73$. There were no interactions of group x valence ($p > 0.1$) or Cohort x group x valence ($p > 0.1$). There was a main effect of group on valence rating $F(2,79)=7.25$, $p=0.009$, $\sim\eta^2=0.083$. For negative words only, there were main effects of both group and Cohort. Depressed subjects rated the negative words as more negative compared to controls' ratings $F(2,79)=7.46$, $p=0.008$, $\sim\eta^2=0.09$. and Cohort 2 rated negative words as more negative $F(2,79)=9.88$, $p=0.002$, eta^2=0.11.

VI-B. Reaction Time preparation. Harmonic means of reaction times (RTs) were calculated within subjects for each condition (28). Outliers (>Tukey Hinges (25^{th} or 75^{th} percentiles) +/-1.5IQR) were Winsorized (rescaled to Hinges+/-1.5IQR).

VI-C. Reaction Time Effects. There was a main effect of word valence on reaction time $F(2,79)=12.11$, $p < .0001$, $\sim\eta^2=.24$. Participants were significantly slower to rate the neutral words compared to the positive words $t(166)=-2.24$, $p=.026$, $d=0.33$ but not consistently slower rating neutral words compared to negative words ($p > .1$) or rating negative words compared to positive ($p>.1$). No significant effects interactions of Cohort ($p = .096$) or group ($p = .088$). In addition, no valence x group x Cohort interaction was present ($p > .1$).

Depressed subjects were slower to respond to all words, Group main effect, $F(2,79) = 7.93$, $p = 0.006$, $\sim\eta^2=.09$ with no Cohort x group interaction $F(2,79) = 4.10$, $p = 0.46$, $\sim\eta^2=.05$. In Cohort 1, controls and depressed participants were not significantly different in their overall reaction times ($p >.1$) however in Cohort 2, depressed participants were significantly slower in overall reaction times compared to controls $t(153)=5.67$, $p>.0001$, $d = 0.94$. Similarly, controls in Cohort 1 were significant slower than controls in Cohort 2 $t(103)=3.30$, $p=.001$, $d=0.65$. Within the depressed group, Cohort 2 was slower than Cohort 1 $t(144.75)=-2.05$, $p=.042$, $d=0.30$.

Importantly, significant contrasts reflect differences in reaction times of no more than 200ms, which is small compared to the smoothing kernel of the hemodynamic response. As such, these differences were not formally accounted for in fMRI analyses, as they would have made virtually no difference.

VI-C. Affect before and after the task. Participants rated their sad, happy, and anxious affect from 1 (not at all) to 5 (very much) before and after the task. Before the task, compared to control participants, depressed participants were more sad, $t(76)=4.86$, $p<.0005$, $D=1.34$ points, $d=1.32$, less happy, $t(76)=3.16$, $p=.002$, $D=-.775$, $d=.73$, and more anxious $t(76)=3.01$, $p=.004$, $D=.85$, $d=.70$. The same differences were present after the task. Paired t-tests on pre/post task affect ratings suggested that among depressed participants, the task appeared to diminish sad, $t(43)=2.46$, $p=.02$, and anxious, $t(43)=2.1$, $p=.04$, and happy affect, $t(43)=2.56$, $p=.01$. Controls' affect did not change across the task ($p's > .18$).

Table S1: Pre-treatment behavioral data

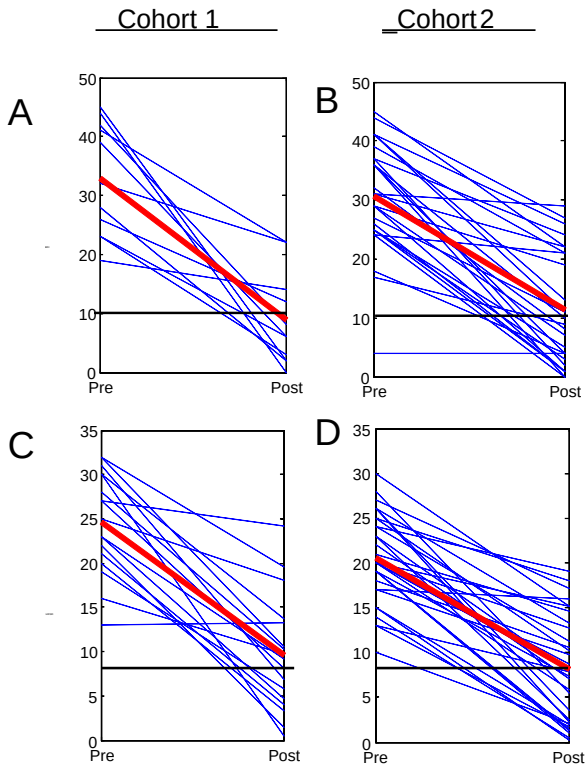
Measure	Depressed Cohort 1 CT	Depressed Cohort 2 CT	Control Cohort 1	Control Cohort 2
Post-scan Emotion ratings of employed words (1= very negative and 7 = very positive)				
Positive words M(SD)	5.22 (0.41)	5.49 (0.70)	5.24 (0.40)	5.66 (0.54)
Negative words M(SD)	2.55 (0.66)	2.11 (0.62)	2.99 (0.65)	2.49 (0.72)
Neutral words M(SD)	4.25 (0.44)	4.05 (0.60)	4.21 (0.48)	4.31 (0.34)
Reaction times on the PRRT in ms M(SD) from 1 (not at all) to 5 (very much)				
Positive words	1413.67 (245.42)	1461.01 (430.98)	1332.68 (375.44)	1096.03 (295.45)
Negative words	1403.82 (269.41)	1637.34 (526.16)	1411.31 (390.26)	1242.20 (374.90)
Neutral words	1588.00 (353.05)	1687.28 (615.15)	1450.21 (432.75)	1154.67 (316.96)
Affect Ratings M(SD)				
Pre-task Sad	2.62(1.02)	2.54(1.20)	1,67 (1.07)	1.26(.56)
Post-task Sad	2.27(.88)	2.18(1.06)	1.58(.900)	1.29(.41)
Pre-task Happy	2.06(1.06)	2.04(1.00)	3,00(1.04)	2.74(1.20)
Post-task Happy	1.47(.64)	1.82(1.02)	3.25(.87)	2.95(1.15)
Pre-task Anxious	3.19(1.28)	2.79(1.10)	2,.17(1.59)	2.00(1.20)
Post-task Anxious	3.07(.96)	2.50(1.14)	2,00(1.13)	1.65(.81)

Author Material-VII. Trajectories of Clinical Change

As shown in Figure S1, all participants decreased in symptomatology on both the BDI and HRSD.

Figure S1: Response to CT by Cohort and Outcome Measure

Figure S1: Response to CT by Cohort and outcome measure. A) Cohort 1, BDI. B) Cohort 2, BDI. C. Cohort 1 HRSD. D. Cohort 2 HRSD.



Author Material-VIII. More complete neuroimaging results tables

A. Table S2a: Table 2 from the primary manuscript, augmented with all relevant statistics and p-values along with statistics from ROC analyses.

ROC statistics include ROC Area Under the Curve (AUC; fair>.7, good>.8, excellent>.9) and a Z-test of its significance.

Cohort	Measure	Residual	Response Prediction (% correct, permutation p, ROC AUC, Z_{AUC}, p_{AUC})	Remission prediction (% correct, permutation p, ROC AUC, Z_{AUC}, p_{AUC})
A. Prediction using sgACC with each Cohort as a sample of interest				
1 Completers (as in (5)) (N=14)	BDI	R^2 = 0.414, F(1,12)=7.781, p=0.018, M(points off)=6.373	84% correct, Threshold = 0.05% change, p=.1 (.04 based on d'), Sensitivity = 100%, Specificity = 86%, d'=2.2	84% correct, Threshold = 0.05% change, p=.08 (.04 based on d') Sensitivity = 100%, Specificity = 75%, d'=2.1
1 Completers and Noncompleters (estimated) (N=16)	BDI	R^2 = 0.358, F(1,15)=7.79, p=0.014, M(points off)=7.045	81% correct, Threshold = 0.05% change, p=.07 Sensitivity = 83%, Specificity = 80%, d'=1.8, ROC AUC=.83, Z_{AUC}=3,2, p_{AUC}<.001	75% correct, Threshold = -0.06% change, p=.14 Sensitivity = 100%, Specificity = 50%, d'=1.4, ROC AUC=.78, Z_{AUC}=2.35, p_{AUC}=.009
(N=17)	HRSD	R^2 = 0.341, F(1,16)=7.772, p=0.014	76% correct, Threshold = 0.08% change, p=.41 Sensitivity = 80%, Specificity = 67%, d'=1.3, ROC AUC=.71, Z_{AUC}=1.6, p_{AUC}=.049	82% correct, Threshold = -.06% change, p=.045 Sensitivity = 100%, Specificity = 57%, d'=1.7, ROC AUC=.85, Z_{AUC}=3,5, p_{AUC}<.001
2 (N=27)	BDI	R^2 = 0.167, F(1,26)=4.91, p=0.04, M(points off)=6.43	78% correct, Threshold = 0.07% change, p=.04 Sensitivity = 40%, Specificity = 100%, d'=1.5, ROC AUC=.60, Z_{AUC}=,88, p_{AUC}=.19	74% correct, Threshold = 0.02% change, p=.04 Sensitivity = 54%, Specificity = 93%, d'=1.6, ROC AUC=.70, Z_{AUC}=2.07, p_{AUC}=.02
(N=32)	HRSD	R^2 = 0.040, F(1,31)=1.244, p=0.273	78% correct, Threshold = 0.08% change, p=.02 Sensitivity = 42%, Specificity = 100%, d'=1.61, ROC AUC=.41, Z_{AUC}=-,82, p_{AUC}=.79	66% correct, Threshold = 0.01% change, p=.38 Sensitivity = 32%, Specificity = 100%, d'=1.32, ROC AUC=.65, Z_{AUC}=1.49, p_{AUC}=.06
1 & 2 combined (N=43 with pre/post	BDI	R^2 = 0.288, F(1,42)=16.61, p<0.0005	79% correct, Threshold = 0.08% change, p=.003,	72% correct, Threshold = 0.02% change, p=.01,

BDI scores (of N=49 in the combined cohort))				Sensitivity = 50%, Specificity = 96%, d'=1.79, ROC AUC=.70, Z_{AUC}=2.6, p_{AUC}=.01	Sensitivity = 38%, Specificity = 95%, d'=1.39, ROC AUC=.75, Z_{AUC}=3.3, p_{AUC}<.001
(N=49)	HRSD	R^2 = 0.12, F(1,48)=6.29, p=0.02		78% correct, Threshold = 0.08% change, p=.001 Sensitivity = 47%, Specificity = 93.8%, d'=1.46, ROC AUC=.37, Z_{AUC}=-1.5, p_{AUC}=.93	69% correct, Threshold = -.06% change, p=.06 Sensitivity = 34%, Specificity = 100%, d'=1.58, ROC AUC=.73, Z_{AUC}=3.06, p_{AUC}=.001
B. Prediction of residual symptoms using a priori network with combined cohort as a sample of interest (N=43)*					
Zero order relationships	BDI	sg ACC	R^2 = 0.29, F(1,42)=16.61, p<0.005		
		R Amygdala	R^2 = 0.16, F(1,42)=7.54, p=0.01		
		L DLPFC	R^2 = 0.20, F(1,42)=10.35, p=0.002		
		BA24 in the VMPFC	R^2 = 0.11, F(1,42)=5.14, p=0.03		
Multivariate relationships, Full model R^2=.43, F(1,38)=7.39, p=.0001	BDI	Constant	β=-1.79, stβ=0, t=1.25, p=.21		
		sgACC	β=43.1, stβ=.47, t=3.79, p=.0005		
		R Amygdala	β=17.2, stβ=.24, t=1.67, p=.10		
		L DLPFC	β=35.65, stβ=.30, t=1.98, p=.05		
		BA24 in the VMPFC	β=-11.98, stβ=-.10, t=.66, p=.52		
C. Generalizable classification from just sgACC or all 4 regions and their connectivity using random forest methodology with Cohort 1 as the training set and Cohort 2 as the test set.**					
1, CT – training set – all brain regions	BDI	R^2 = 0.28, p<0.02, M(points off)=7.5		75% correct, Sensitivity = 75%, Specificity = 90%, d'=1.28, p=.023,	69% correct, Sensitivity = 62%, Specificity = 75%, d'=.99, p=.057, ROC

			ROC AUC=.88***, Z_{AUC}=3.86, p_{AUC}<.0005	AUC=.76, Z_{AUC}=2.17, p_{AUC}=.01
	HRSD	R^2 = 0.21, p<0.01, M(points off)=4.5	76% correct, Sensitivity = 92%, Specificity = 40%, d'=1.13, p=.028, ROC AUC=.83, Z_{AUC}=2.7, p_{AUC}=.004	65% correct, Sensitivity = 57%, Specificity = 70%, d'=.7, p=.12, ROC AUC=.72, Z_{AUC}=1.84, p_{AUC}=.03
2, CT – Generalization set – initial severity alone	BDI	R^2 = 0, p<0.74, M(points off)=8.9	48% correct, Sensitivity = 53%, Specificity = 40%, d'=-.18, p=.63, ROC AUC=.66, Z_{AUC}=.913, p_{AUC}=.181	38% correct, Sensitivity = 23%, Specificity = 54%, d'=-.64, p=.92, ROC AUC=.48, Z_{AUC}=-.09, p_{AUC}=.54
	HRSD	R^2 = 0, p=0.63, M(points off)=5.4	56% correct, Sensitivity = 60%, Specificity = 50%, d'=.25, p=.29, ROC AUC=.7, Z_{AUC}=2.01, p_{AUC}=.02	47% correct, Sensitivity = 38%, Specificity = 53%, d'=-.23, p=.7, ROC AUC=.51, Z_{AUC}=.173, p_{AUC}=.43
2, CT - Generalization set – initial severity + sgACC	BDI	R^2 = 0, p=0.07, M(points off)=6.7	74% correct, Sensitivity = 100%, Specificity = 30%, d'=1.04, p=.03, ROC AUC=.78, Z_{AUC}=2.96, p_{AUC}=.002	78% correct, Sensitivity = 86%, Specificity = 69%, d'=1.57, p=.003, ROC AUC=.75, Z_{AUC}=2.72, p_{AUC}=.003
	HRSD	R^2 = 0.02, p=.07, M(points off)=3.9	81% correct, Sensitivity = 95%, Specificity = 58%, d'=1.86, p<.001, ROC AUC=.83, Z_{AUC}=3.95, p_{AUC}=.0004	62% correct, Sensitivity = 69%, Specificity = 58%, d'=.7, p=.08, ROC AUC=.68, Z_{AUC}=2.0, p_{AUC}=.02
2, CT – Generalization set – initial severity + all brain regions (retained: Baseline BDI, LDLPFC, sgACC, phi_LDLPFC_sgACC)	BDI	R^2 = 0.15, p<0.003, M(points off)=5.6	81% correct, Sensitivity = 100%, Specificity = 44%, d'=1.43, p=.0005, ROC AUC=.81, Z_{AUC}=3.36, p_{AUC}=.0004	63% correct, Sensitivity = 79%, Specificity = 46%, d'=.7, p=.086, ROC AUC=.69, Z_{AUC}=1.92, p_{AUC}=.03
	HRSD	R^2 = 0.07, p=.02, M(points off)=3.8	75% correct, Sensitivity = 95%, Specificity = 42%, d'=1.43, p=.002, ROC AUC=.85, Z_{AUC}=4.54, p_{AUC}<.00005	66% correct, Sensitivity = 85%, Specificity = 53%, d'=1.09, p=.015, ROC AUC=.76, Z_{AUC}=3.1, p_{AUC}=.001

* Classification not evaluated in the multivariate model without robust estimation (2B) given the potential for type I error – rather, evaluations in 2C reflect robust estimations from the multivariate model.

**p-values for random forests were estimated via permutation tests on %-correct using the same random-forest methodology. Regression and classification accuracy estimates unbiased computed on out-of-bag samples for robust generalization.
*** Multivariate ROC AUC values in C. were obtained using a loss function that differentially weighted sensitivity and specificity.

B. Association of Z-scores with response defined as clinical change

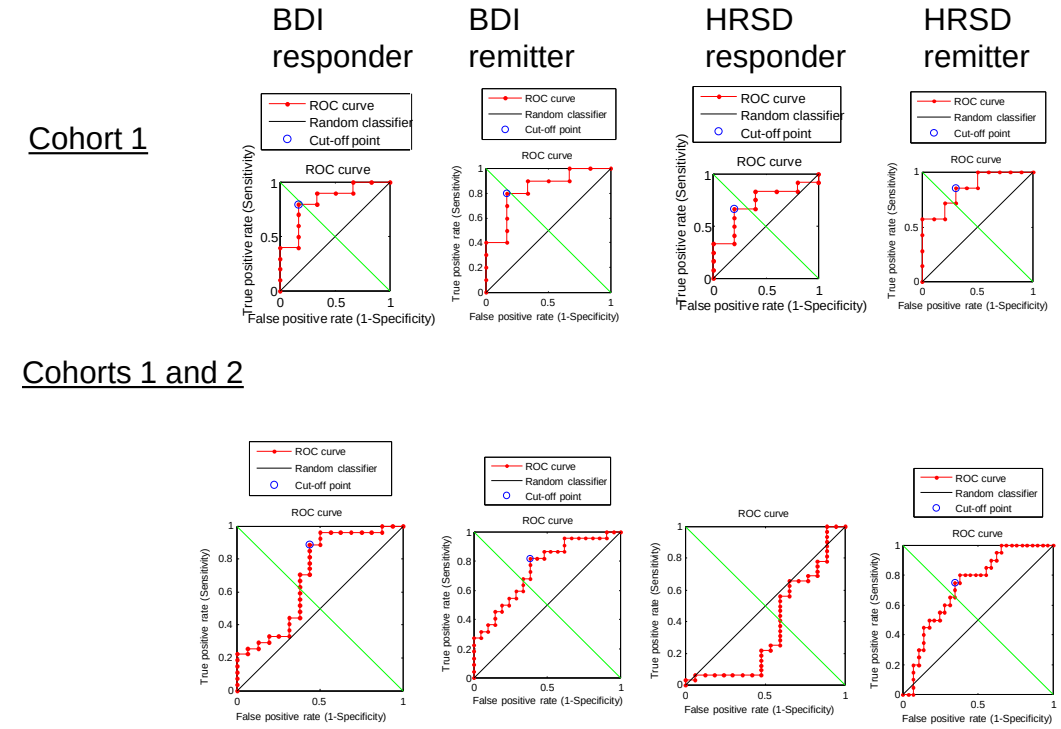
As shown in Table S2B, clinical change (post-pre BDI or HRSD) was strongly associated with change in standardized (Z) scores computed relative to the mean and standard deviation of control participants.

Table S2B: Association of anatomical BA25 activity with change in severity using Z scores

Cohort, Intervention	Measure	Change	Response Prediction (% correct, sensitivity, specificity, X^2, p)	Remission prediction (% correct, sensitivity, specificity, X^2, p)
1, CT	BDI	R^2 = 0.54, F(1,15)=16.32, p=0.001, M(points off)=6.4	81% correct, Threshold Z = .47, p=.07	75% correct, Threshold Z=-.52, p=.14
	HRSD	R^2 = 0.225, F(1,16)=4.367, p=0.054, M(points off)=5.167	76% correct, Threshold Z = 0.75, p=.40	82% correct, Threshold Z = -.5, p=.03
2, CT	BDI	R^2 = 0.12, F(1,25)=3.24, p=0.08, M(points off)=7.17	78% correct, Threshold Z = .68, p=.03	74% correct, Threshold Z = 0.22 p=.04
	HRSD	R^2 = 0.062, F(1,31)=1.968, p=0.171, M(points off)=4.278	78% correct, Threshold Z = 0.74, p=.02	66% correct, Threshold Z = 0.08, p=.36
1 & 2 combined CT	BDI	R^2 = 0.288, F(1,42)=16.605, p=0.000, M(points off)=7.172	78% correct, Threshold Z = .68, p=.04	74% correct, Threshold Z= .22, p=.045
	HRSD	R^2 = 0.089, F(1,48)=4.572, p=0.038, M(points off)=4.900	78% correct, Threshold Z = 0.74, p=.02	66% correct, Threshold Z = .07, p=.37

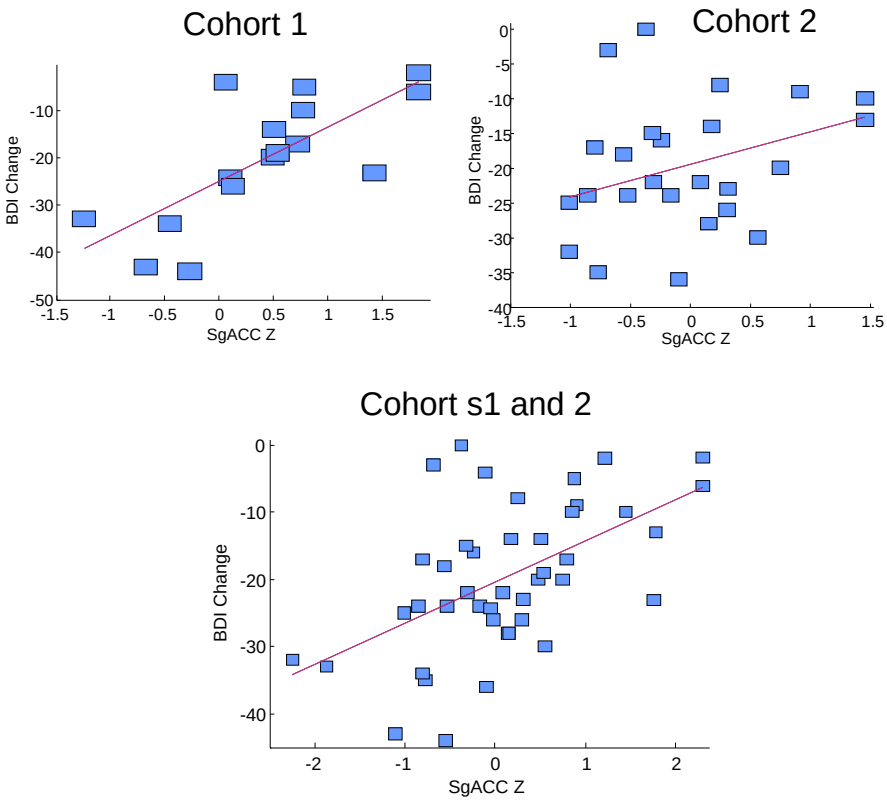
Author Material-IX. ROC curves for response and remission computed using BA25

Figure S2: ROC curves for response and remission computed using BA25



Author Material-X. Association of Z scores with changes in symptoms

Figure S3. Association of sgACC Z-score with #points change in BDI-II score

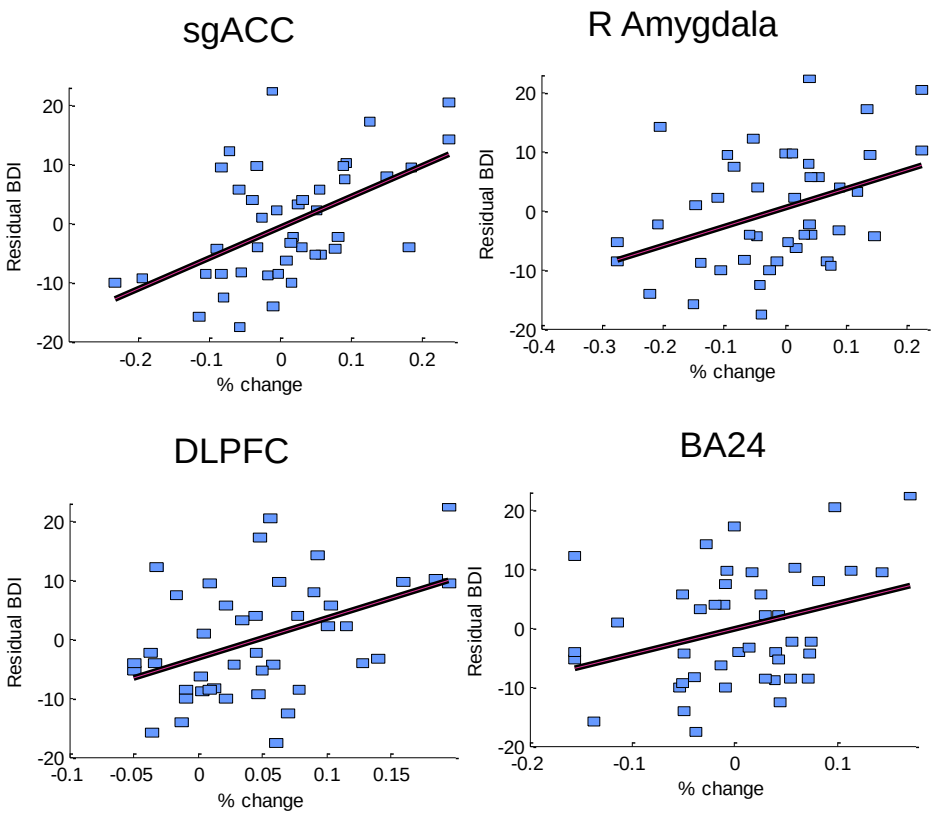


Author Material-XI. Associations of multiple regions and measures with response

XI-A. Univariate Associations.

Figure S4 shows the scatterplot of bivariate associations with each of the proposed regions with residual BDI severity across the cohorts (the combined sample was used to preserve adequate statistical power given the use of multiple predictors).

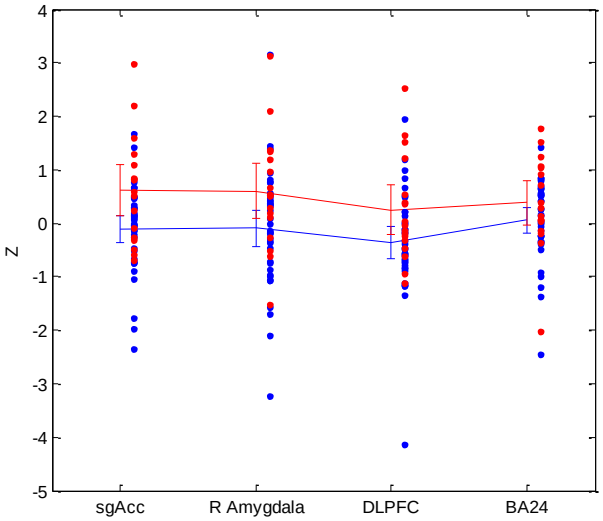
Figure S4: Association of activity in multiple a priori regions with response in the combined sample



XI-B. Profiles for Responders and Non-Responders

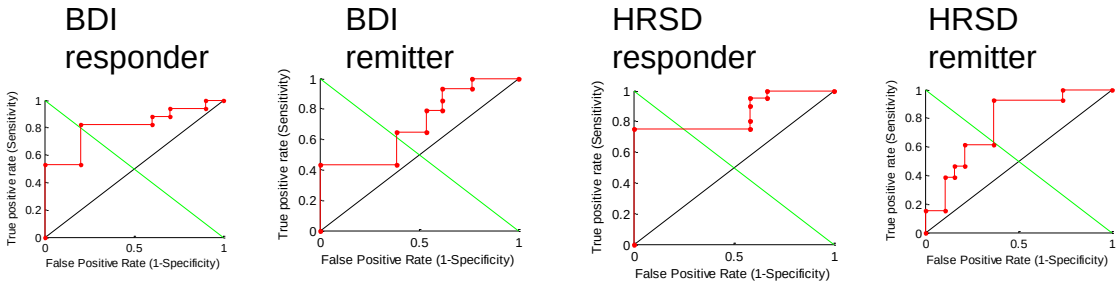
To aid clinical interpretation at a single subject level, Figure S5 shows the profile of activities for BDI responders and non-responders across regions expressed as Z-scores of control participants. Thus, participants who have profiles that fall below the line of a detectable positive response to stimuli on the task are likely to respond to CT. In contrast, participants who do have a detectable regulatory response are less likely to be helped by CT.

Figure S5: Profiles for responders and non-responders across multiple anatomically defined regions



XI-C. ROC Curves – Multivariate tests accounting for multiple regions

Figure S6. ROC Curves for the Generalization set from the full model (Baseline BDI, retained regions: LDLPFC, BA25, phi_LDLPFC_BA25).



XI-D. Other likely-candidates: pre-treatment severity, rumination, pupil, demographics

To examine the extent to which fMRI predictions were 1) augmented by and 2) accounted for by more traditional and inexpensive measures we examined a hierarchical regression on residual BDI scores in the combined sample. This was particularly important as the sgACC region had relatively low sensitivity when entered alone. Here, scanner was entered on the first step. sgACC was entered on the second step. Self-reported rumination (rumination scale from the Response Styles Questionnaire (29), administered within two weeks of the scan, was entered on the fourth step. We have previously shown rumination to predict response to CT (30). Variables associated with pupillary motility in response to negative words acquired during the fMRI scan, were also entered on the third step. Pupil dilation yields a physiological measure of cognitive and affective processing (31, 32), is associated with prefrontal control (33) and limbic activity (34) and, in association with another task and sample, predicted response and remission in CT (35). Pupillary response variables calculated in response to negative words, included average response amplitude, peak response amplitude, and average response during the last seconds of the trial. On the fourth step, traditional demographics including age, education, and gender were entered. A second analysis entering pre-treatment severity on the second step was also examined.

No step other than step two on which the sgACC was entered significantly predicted additional variance in residual BDI scores or increased the adjusted R^2, and at no step was any entered variable significantly predictive except for sgACC which was significantly predictive at every step as shown in Table S3.

Table S3: Additional variance explained by non-fMRI predictors

Step	Variables entered	Tot R^2 (Adj R^2)	ΔR^2	Statistic	Standardized B, p for sgACC	Max standardized B for any other variable
1	Scanner	.02 (-.03)	.02	$F\Delta$ (1,22)=.3, p=.57	N/A	B=.12, p=.57
2	sgACC	.38 (.32)	.36	$F\Delta$ (1,21)=12.5, p=.002	B=.63, p=.002	B=.27, p=.14
3	Rumination, pupil	.46 (.26)	.07	$F\Delta$ (4,17)=.6, p=.67	B=.74, p=.005	B=.26, p=.25
4	Demographics	.57 (.31)	.11	$F\Delta$ (3,14)=1.3, p=.30	B=.78, p=.01	B=.32, p=.09

When demographic variables (age, education, gender) were entered on the first step, they did not account for significant variance, F(3,20)=.93, p=.45, Tot R^2 (Adj R^2) =.12(-0.01).

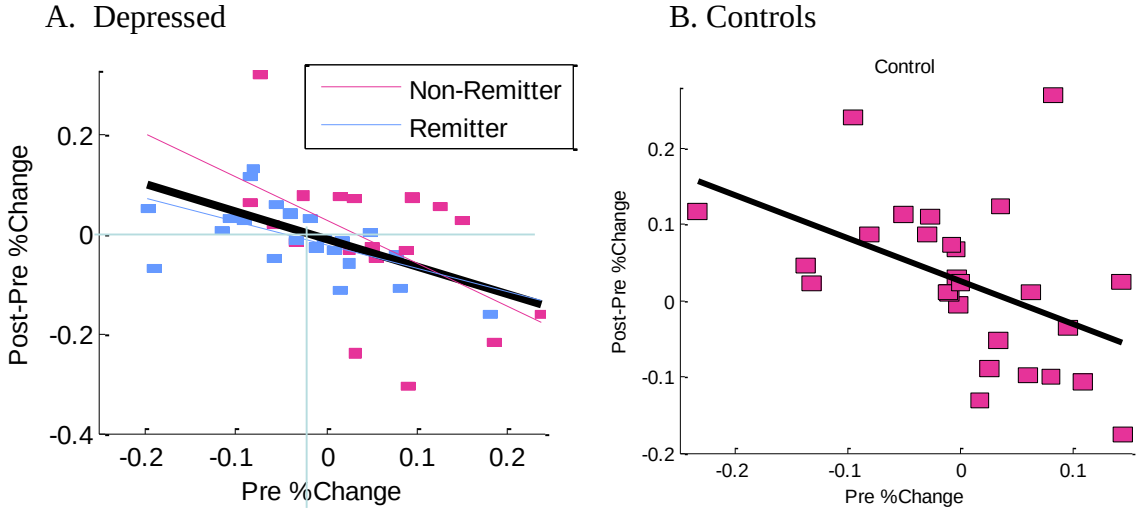
To preserve as parallel a process as possible a strategy for examining response, logistic regressions following the same strategy were examined, also including initial severity as a predictor. Scanner explained virtually no variance, with all participants predicted to recover, 65% correct classification, $X^2(1)=.34$, p=.6, Cox & Snell R^2 =.01. sgACC entered on the second step explained significantly more variance in recovery, 74% correct classification, $X^2(1)=.5.83$, p=.02, Cox & Snell R^2 =.15. Demographic variables and initial severity entered on the third step did not increase explained variance, 76% correct classification, $X^2(4)=.6.4$, p=.17, Cox & Snell R^2 =.28; N=3 participants changed classification. When rumination and pupil variables were entered into the model only 24 cases were available. Thus sgACC was no longer a significant response predictor above and beyond scanner, N=58% correct, $X^2(1)=3.6$, p=.06, Cox & Snell R^2 =.15. Rumination and pupil improved classification nonsignificantly, N=71% correct, $X^2(4)=4.5$, p=.34, Cox & Snell R^2 =.29, with demographic variables not significantly increasing classification due to low power, N=75% correct, $X^2(4)=3.8$, p=.43, Cox & Snell R^2 =.40. With the exception of BA25 (p=.097), no variables in the final model approached significance p>.16.

When demographic variables (age, education, gender) were entered on the first step, they did not account for significant variance, 68% correct classification, $X^2(3)=5.08$, p=.17, Cox & Snell R^2 =.13. Pretreatment severity explained no more variance, 68% correct classification, $X^2(1)=.06$, p=.8, Cox & Snell R^2 =.13.

Author Material-XII. Associations of pre-treatment sgACC activity with sgACC change

XII-A. Continuous change in sgACC. As shown in Figure S7A, depressed participants with the lowest pre-treatment $sgACC_Z$ displayed relatively little change in sgACC activity whereas participants with high sgACC activity generally decreased yielding a significant relationship between pre-treatment activity and change R^2=0.259, F(1,39)=13.310, p=0.001. As shown in Figure S7B, this could represent regression to the mean as controls with the lowest $sgACC_Z$ also displayed the strongest increases, R^2=0.161, F(1,26)=4.791, p=0.038, with no group x pre-treatment interaction, R^2<.01.

Figure S7. Relationship between pre-treatment activity and change



XII-B. Effects of treatment on sgACC activity in remitters with low pre-treatment activity.

To examine treatment effects, versus regression to the mean, we considered whether sgACC activity increased among more depressed remitters who had sgACC activity below the threshold for predicting remission (0.02% signal change) than controls below that threshold and depressed non-remitters below the threshold tested after treatment (after 16 weeks for controls). As shown in Table S4, depressed remitters with low pre-treatment activity had a non-significantly smaller proportion of participants who increased and lower mean level of increase than did either controls of non-remitters with low pre-treatment activity. Thus, we cannot conclude that sgACC activity increased as a function of treatment rather than regression to the mean.

Table S4. Change in sgACC activity in depressed participants with pre-treatment sgACC activity lower than the threshold for predicting remission compared to comparably low controls and depressed non-remitters

	N increased	Fisher's exact test v. depressed low remitter, p	Mean %-change increase	t-test v. Depressed low remitter	t-test pre v. post
Depressed low remitter (N=17)	11 (64%)		.018%		t(16)=1.13, p=.27, d=.27
Control low (N=16)	14 (87%)	.13	.06%	t(31)=-1.66, p=.11, d=-.57	t(15)=2.89, p=.01, d=.72
Depressed low non-remitter (N=6)	5 (83%)	.38	.09%	t(21)=-1.97, p=.06, d=-.93	t(5)=2.32, p=.06, d=.94

XII-C. Improved classification using post-treatment. Using grid-search, $sgACC_{Post}$ improved classification of remission from d'=1.4 (95% sensitivity, 39% specificity) to d'=1.8 (95% sensitivity, 56% specificity) increasing specificity as non-remitters often had relatively high $sgACC_{Z\text{-}Pre}$ (Z>-1.01) which increased, relatively, in treatment

(Z_{post} -Z_{pre}>-.59).

Author Material XIII. Positive and Neutral words

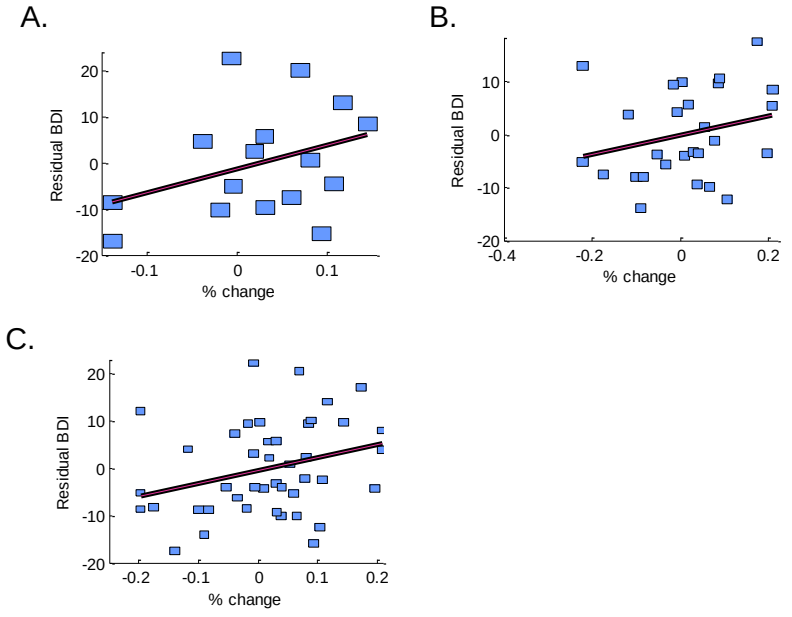
As the primary purpose of this manuscript involved a replication of (5), which examined only negative words, primary analyses involved examination of negative words. That said, the paradigm allowed us to investigate the specificity of prediction to just negative words. Prediction from positive and neutral words could suggest that prediction regards self-referent processing in general rather than simple rumination on presented negative information.

In the full sample with residual BDI values (N=43), associations of residual symptomatology with negative words ($r=.538$, $p<.005$) were significant, unlike associations with positive words ($r=.13$, $p=.39$) and neutral words ($r=.27$, $p=.07$). Together, positive and neutral words explained 8.4% of the variation in residual symptoms, $F(2,40)=1.8$, $p=.17$. When added to that model, negative words explained an additional 21.6% of the variation, $\Delta F(1,39)=12.0$, $p=.001$. Thus, associations of response with sgACC activity appear specific to negative words.

Qualitatively similar results were apparent for just Cohort 1 (negative words added 19% of the variation to the 23% of variation explained by positive and neutral words) and Cohort 2 (negative words added 17% of variation to the 6% explained by positive and neutral words).

Figure S8 shows graphically that, associations of residual symptomatology with responses to positive words at scans 4-7 (analog of primary analyses for negative words in Figure 2) were small in Cohort 1 (Figure S8A), Rsq = 0.122, $F(1,15)=1.952$, $p=0.184$, and Cohort 2 (Figure S8b), Rsq = 0.064, $F(1,26)=1.721$, $p=0.20$, and thus small in the combined cohort (Figure S8c), Rsq = 0.086, $F(1,42)=3.844$, $p=0.05$. Associations with neutral words were even smaller, combined cohort Rsq = 0.062, $F(1,26)=1.652$, $p=0.21$.

Figure S8: Anatomically defined sgACC activity in response to positive words was weakly correlated with stronger clinical response (decreased residual BDI) in A) Cohort 1, B) Cohort 2, and C) the combined cohort.



References

1. Beck AT, Rush AJ, Shaw BF, Emery G. Cognitive therapy of depression. New York: Guilford Press; 1979.
2. Beck JS. Questions and Answers about Cognitive Therapy: http://www.beckinstitute.org/InfoID/220/RedirectPath/Add1/FolderID/237/SessionID/%7B2597457D-2029-46FB-B757-639DD4D90CB4%7D/InfoGroup/Main/InfoType/Article/PageVars/Library/InfoManage/Zoom.htm.
3. Greenberger D, Padesky CA. Mind Over Mood New York Guilford; 1995.
4. Young J, Beck AT. Cognitive THerapy Scale Rating Manual. 1980.
5. Siegle GJ, Carter CS, Thase ME. Use of fMRI to predict recovery from unipolar depression with Cognitive Behavior Therapy. Am J Psychiatry. 2006;163(4):735-8.
6. Bandelow B, Baldwin DS, Dolberg OT, Andersen HF, Stein DJ. What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder? J Clin Psychiatry. 2006;67(9):1428-34.
7. Keller MB. Past, present, and future directions for defining optimal treatment outcome in depression: remission and beyond. JAMA. 2003;289(23):3152-60.
8. Riedel M, Moller HJ, Obermeier M, Schennach-Wolff R, Bauer M, Adli M, et al. Response and remission criteria in major depression - A validation of current practice. J Psychiatr Res. in press.
9. Siegle GJ. The Balanced Affective Word List Creation Program. Available Web at http://www.sci.sdsu.edu/CAL/wordlist/; 1994.
10. Bradley MM, Lang PJ. Affective Norms for English Words ANEW Technical Manual and Affective Ratings Gainsville FL The Center for Research in Psychophysiology University of Florida; 1997.
11. Cox R. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers in Biomedical Research. 1996;29:162-73.
12. Holmes C, Hoge R, Collins L, Woods R, Toga A, Evans A. Enhancement of MR images using registration for signal averaging. Journal of Computer Assisted Tomography. 1998;22(2):324-33.
13. Woods RP, Mazziotta JC, Cherry SR. MRI PET registration with automated algorithm. Journal of Computer Assisted Tomography. 1993;17:536-46.
14. Siegle GJ, Steinhauer SR, Thase ME, Stenger VA, Carter CS. Can't shake that feeling: fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. Biological Psychiatry. 2002;51:693-707.
15. Siegle GJ, Thompson W, Carter CS, Steinhauer SR, Thase ME. Increased amygdala and decreased dorsolateral prefrontal BOLD responses in unipolar depression: Related and independent features. Biol Psychiatry. 2007;61(2):198-209.
16. Woods RP, Grafton ST, Watson JD, Sicotte NL, Mazziotta JC. Automated image registration: II. Intersubject validation of linear and nonlinear models. J Comput Assist Tomogr. 1998;22(1):153-65.
17. Woods RP, Grafton ST, Holmes CJ, Cherry SR, Mazziotta JC. Automated image registration: I. General methods and intrasubject, intramodality validation. J Comput Assist Tomogr. 1998;22(1):139-52.

18. Mayberg HS, Brannan SK, Mahurin RK, Jerabek PA, Brickman JS, Tekell JL. Cingulate function in depression: A potential predictor of treatment response Neuroreport. 1997;8(4):1057-61.
19. Brannan SK, Mayberg HS, McGinnis S, Silva JA, Tekell J, Mahurin RK, et al. Cingulate metabolism predicts treatment response: A replication. Biololgical Psychiatry. 2000;47:107S #355.
20. Brody AL, Saxena S, Mandelkern MA, Fairbanks LA, Ho ML, Baxter LR. Brain metabolic changes associated with symptom factor improvement in major depressive disorder. Biol Psychiatry. 2001;50(3):171-8.
21. Davidson RJ, Irwin W, Anderle MJ, Kalin NH. The neural substrates of affective processing in depressed patients treated with venlafaxine. American Journal of Psychiatry. 2003;160(1):64-75.
22. Chen CH, Ridler K, Suckling J, Williams S, Fu CH, Merlo-Pich E, et al. Brain Imaging Correlates of Depressive Symptom Severity and Predictors of Symptom Improvement After Antidepressant Treatment. Biol Psychiatry. 2007.
23. Keedwell PA, Drapier D, Surguladze S, Giampietro V, Brammer M, Phillips M. Subgenual cingulate and visual cortex responses to sad faces predict clinical outcome during antidepressant treatment for depression. J Affect Disord. 2009.
24. Nitschke JB, Sarinopoulos I, Oathes DJ, Johnstone T, Whalen PJ, Davidson RJ, et al. Anticipatory activation in the amygdala and anterior cingulate in generalized anxiety disorder and prediction of treatment response. Am J Psychiatry. 2009;166(3):302-10.
25. Langenecker SA, Kennedy SE, Guidotti LM, Briceno EM, Own LS, Hooven T, et al. Frontal and limbic activation during inhibitory control predicts treatment response in major depressive disorder. Biol Psychiatry. 2007;62(11):1272-80.
26. Roy M, Harvey PO, Berlim MT, Mamdani F, Beaulieu MM, Turecki G, et al. 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. 2010;35(3):152-62.
27. Pizzagalli D, Pascual-Marqui RD, Nitschke JB, Oakes TR, Larson CL, Abercrombie HC, et al. Anterior cingulate activity as a predictor of degree of treatment response in major depression: evidence from brain electrical tomography analysis. Am J Psychiatry. 2001;158(3):405-15.
28. Ratcliff R. Methods for dealing with reaction time outliers. Psychological Bulletin. 1993;114:510-32.
29. Nolen-Hoeksema S, Morrow J, Fredrickson BL. Response styles and the duration of episodes of depressed mood. Journal of Abnormal Psychology. 1993;102(1):20-8.
30. Jones NP, Siegle GJ, Thase ME. Effects of rumination and initial severity on remission to Cognitive Therapy for depression. Cognitive Therapy and Research. 2008;32(4):591-604.
31. Beatty J. Task-evoked pupillary responses processing load and the structure of processing resources Psychological Bulletin. 1982;91:276-92.
32. Steinhauer SR, Hakerem G. The pupillary response in cognitive psychophysiology and schizophrenia. Annals of the New York Academy of Sciences. 1992;658:182-204.
33. Siegle GJ, Steinhauer SR, Stenger V, Konecky R, Carter CS. Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. NeuroImage. 2003;20(1):114-24.

34. Koikegami H, Yoshida K. Pupillary dilation induced by stimulation of amygdaloid nuclei. Folia Pychiatrica Neurologica Japonica. 1953;7:109-25.
35. Siegle GJ, Steinhauer SR, Friedman ES, Thompson WS, Thase ME. Remission prognosis for cognitive therapy for recurrent depression using the pupil: utility and neural correlates. Biol Psychiatry. 2011;69(8):726-33.