JAMA Psychiatry | Original Investigation | META-ANALYSIS

Altered Brain Activity in Unipolar Depression Revisited
Meta-analyses of Neuroimaging Studies

Veronika I. Müller, PhD; Edna C. Cieslik, PhD; Ilinca Serbanescu, MSc; Angela R. Laird, PhD; Peter T. Fox, MD; Simon B. Eickhoff, MD

IMPORTANCE During the past 20 years, numerous neuroimaging experiments have investigated aberrant brain activation during cognitive and emotional processing in patients with unipolar depression (UD). The results of those investigations, however, vary considerably; moreover, previous meta-analyses also yielded inconsistent findings.

OBJECTIVE To readdress aberrant brain activation in UD as evidenced by neuroimaging experiments on cognitive and/or emotional processing.

DATA SOURCES Neuroimaging experiments published from January 1, 1997, to October 1, 2015, were identified by a literature search of PubMed, Web of Science, and Google Scholar using different combinations of the terms *fMRI (functional magnetic resonance imaging), PET (positron emission tomography), neural, major depression, depression, major depressive disorder, unipolar depression, dysthymia, emotion, emotional, affective, cognitive, task, memory, working memory, inhibition, control, n-back,* and *Stroop*.

STUDY SELECTION Neuroimaging experiments (using fMRI or PET) reporting whole-brain results of group comparisons between adults with UD and healthy control individuals as coordinates in a standard anatomic reference space and using an emotional or/and cognitive challenging task were selected.

DATA EXTRACTION AND SYNTHESIS Coordinates reported to show significant activation differences between UD and healthy controls during emotional or cognitive processing were extracted. By using the revised activation likelihood estimation algorithm, different meta-analyses were calculated.

MAIN OUTCOMES AND MEASURES Meta-analyses tested for brain regions consistently found to show aberrant brain activation in UD compared with controls. Analyses were calculated across all emotional processing experiments, all cognitive processing experiments, positive emotion processing, negative emotion processing, experiments using emotional face stimuli, experiments with a sex discrimination task, and memory processing. All meta-analyses were calculated across experiments independent of reporting an increase or decrease of activity in major depressive disorder. For meta-analyses with a minimum of 17 experiments available, separate analyses were performed for increases and decreases.

RESULTS In total, 57 studies with 99 individual neuroimaging experiments comprising in total 1058 patients were included; 34 of them tested cognitive and 65 emotional processing. Overall analyses across cognitive processing experiments ($P > .29$) and across emotional processing experiments ($P > .47$) revealed no significant results. Similarly, no convergence was found in analyses investigating positive (all $P > .15$), negative (all $P > .76$), or memory (all $P > .48$) processes. Analyses that restricted inclusion of confounds (eg, medication, comorbidity, age) did not change the results.

CONCLUSIONS AND RELEVANCE Inconsistencies exist across individual experiments investigating aberrant brain activity in UD and replication problems across previous neuroimaging meta-analyses. For individual experiments, these inconsistencies may relate to use of uncorrected inference procedures, differences in experimental design and contrasts, or heterogeneous clinical populations; meta-analytically, differences may be attributable to varying inclusion and exclusion criteria or rather liberal statistical inference approaches.

JAMA Psychiatry. 2017;74(1):47-55. doi:10.1001/jamapsychiatry.2016.2783
Published online November 9, 2016.

Invited Commentary page 56

Supplemental content at jamapsychiatry.com

Author Affiliations: Institute of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany (Müller, Cieslik, Serbanescu, Eickhoff); Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany (Müller, Cieslik, Eickhoff); Department of Physics, Florida International University, Miami (Laird); Research Imaging Institute, University of Texas Health Science Center, San Antonio (Fox); Research Service, South Texas Veterans Administration Medical Center, San Antonio (Fox); State Key Laboratory for Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, China (Fox).

Corresponding Author: Veronika I. Müller, PhD, Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Wilhelm-Johnen-Straße, D-52428 Jülich, Germany (v.mueller@fz-juelich.de).

Unipolar depression (UD) is one of the leading causes of disease burden worldwide.[1] In line with the characteristic symptoms of low mood and reduced activity and concentration,[2] patients with UD show impairments of emotional and cognitive processing.[3,4] Numerous neuroimaging experiments[5-8] have investigated the neural correlates underlying these impairments, but their results vary considerably. This variance might be attributable to sampling effects in this heterogeneous disorder but also to the high degree of experimental and analytic flexibility in neuroimaging. Consequently, several quantitative meta-analyses have been performed to delineate brain regions consistently implicated in UD[9-14] (**Figure 1** illustrates the different steps of a meta-analysis). However, these meta-analyses also yielded inconsistent findings (**Figure 2**).

This divergence across meta-analytic findings is perplexing. Several factors contribute to this predicament. First, most previous meta-analyses[9,10,12-14] used the activation likelihood estimation (ALE) approach to determine convergence of findings across experiments. Given that the null distribution in ALE reflects a random spatial association between findings across the entire brain, all included coordinates must be derived from whole-brain analyses. To our knowledge, most classic, explicit region-of-interest (ROI) analyses were not considered in previous UD meta-analyses. In contrast, hidden ROI analyses were often included by way of considering experiments of partial-brain coverage or reporting contrasts that were masked with a main effect in control individuals. However, inclusion of such results in neuroimaging meta-analyses render the null distribution and hence inference inappropriate (eTable 1 in the Supplement shows ROIs included in previous meta-analyses).

Second, most previous meta-analyses attempted to correct for multiple comparisons by controlling the (voxel-level) false discovery rate (FDR), which is invalid for topological inference on smooth data[15,16] and leads to inflated positive findings. Third, some earlier meta-analyses were performed across relatively low numbers of experiments. These meta-analyses have low power and are prone to yield clusters of convergence that are almost exclusively driven by single experiments.[16] Together, these factors may implicate a high number of spurious findings.

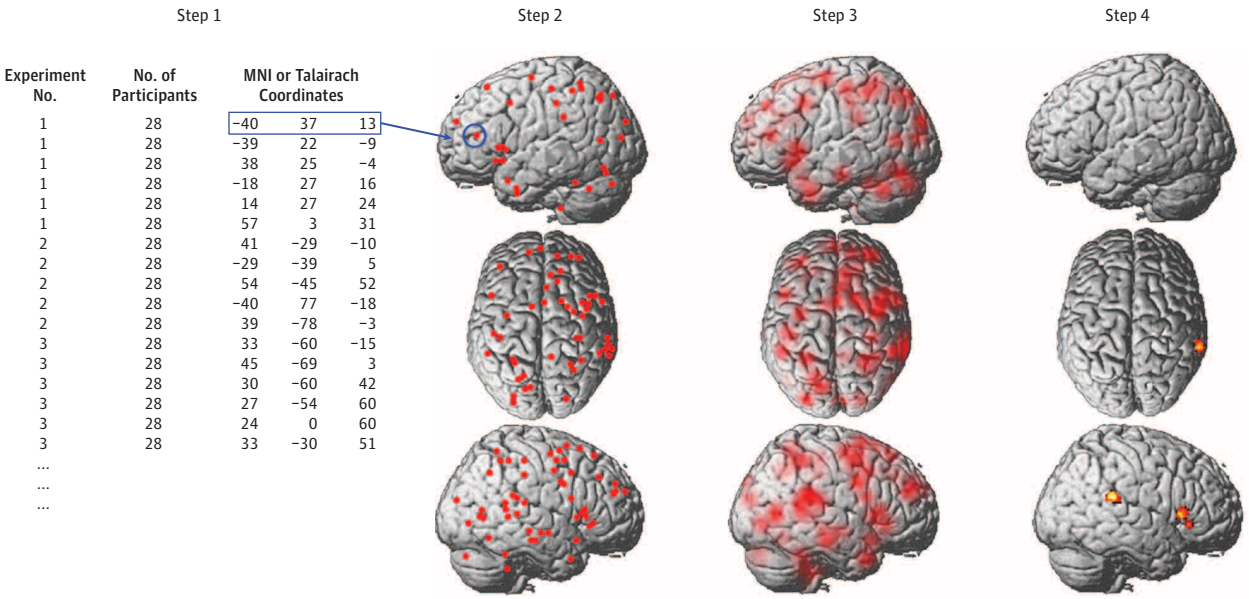
Key Points

Question How consistent are results of experiments investigating aberrant brain activity in unipolar depression and previous meta-analyses testing this topic?

Findings This conceptual replication of meta-analyses of 99 neuroimaging experiments in unipolar depression did not reveal any convergence, which is at odds with the findings of previous meta-analyses.

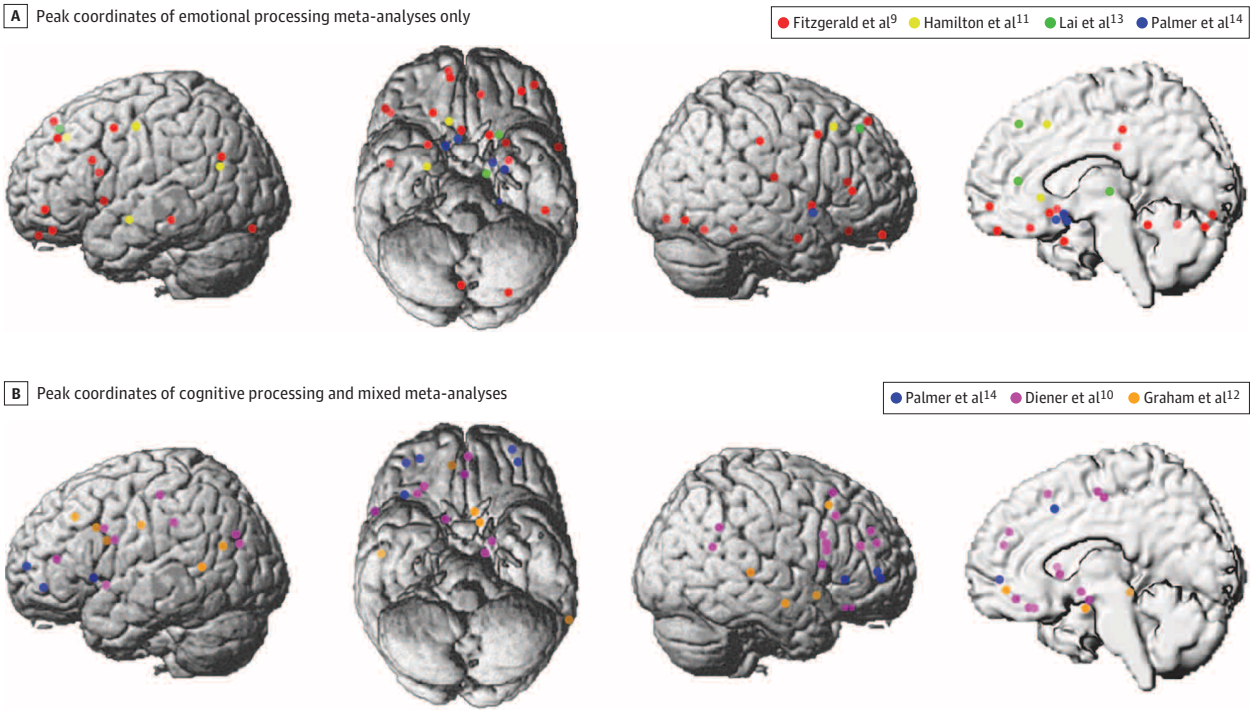
Meaning This result highlights the importance of reproducing previous results to be able to discover real effects for individual neuroimaging studies and for meta-analyses.

Figure 1. Schematic Illustration of the Steps of a Meta-analysis



In step 1, coordinates reported in the experiments included in the analysis are extracted by creating a table with all x, y, and z coordinates in either Montreal Neurological Institute (MNI) or Talairach space. In this example, the coordinates of more than 3 experiments were included indicated by the ellipses at the end of the example table. In step 2, these coordinates are transformed into the same coordinate space and projected on a brain template for display. For example, the first coordinate reported in experiment 1 is projected on a brain template based on its x, y, and z coordinates (marked in blue). In step 3, the spatial uncertainty associated with each coordinate is accounted for by modeling gaussian probability distributions around each coordinate. In step 4, the resulting activation likelihood estimation scores are compared with a null distribution reflecting a random spatial association between experiments, and results are thresholded and corrected for multiple comparisons.

Figure 2. Illustration of Reported Peak Coordinates of Convergence Found in Previous Meta-analyses Investigating Aberrant Brain Activation in Unipolar Depression



Mixed processing indicates analyses across cognitive and emotional experiments. Each color represents a separate meta-analysis. Hardly any overlap is seen between colors, showing that results of previous meta-analyses investigating a similar topic are inconsistent. A, Previous meta-analyses of emotional processing experiments only include Fitzgerald et al,[9] Hamilton et al,[11] Lai et al,[13] and Palmer et al.[14] B, Previous meta-analyses of cognitive processing experiments includes Palmer et al[14]; of mixed processing, Diener et al[10] and Graham et al.[12]

Finally, the focus of investigation has been variable across meta-analyses, with some summarizing cognitive and emotional experiments, others focusing on emotional or cognitive ones only, and yet others focusing on more specific aspects. Although in itself perfectly reasonable, this heterogeneity combined with the bias introduced by including ROI analyses and the high likelihood of false-positive findings may explain the heterogeneity seen in Figure 2.

Thus, the objective of the present investigation is to readdress aberrant brain activation during cognitive and emotional processing in UD using strict quality control and state-of-the-art meta-analyses. Furthermore, we aim to evaluate whether the lack of replication of meta-analyses in UD is the result of methodological problems or to differences in the specific focus of investigations. We thus performed analyses comparable to all objectives of previous meta-analyses. The present work is not a reproduction of previous meta-analyses (ie, including the same experiments as well as the same analytic procedures) but rather a conceptual replication of them. If inconsistent results of previous meta-analyses are attributable to differences in the specific focus of investigation, we should be able to replicate previous results with the respective meta-analysis focusing on the same aspect. A failure of replication, however, would point to methodological problems.

Methods

Inclusion and Exclusion Criteria

Details on the literature research can be found in the eMethods in the Supplement. In brief, neuroimaging experiments published from January 1, 1997, to October 1, 2015, were identified by a literature search of PubMed, Web of Science, and Google Scholar. To provide comprehensive, best-practice analyses of aberrant activation in patients with UD, we applied inclusion and exclusion criteria discussed below.

Criteria Related to the Investigation Participants

Included experiments statistically contrasted neural activation between an adult (>18 years) UD group (based on *DSM-IV-TR*[17] and *DSM-5*[18] or *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*[2]) and a group of healthy controls. Experiments investigating patients with comorbidities were included, but with the criterion of UD as the main diagnosis. Exclusion of comorbidity would have reduced the size of the experiments by half. Furthermore, we only included experiments investigating patients with UD and current depressive clinical symptoms, whereas those experiments that investigated effects in groups with remission of symptoms were excluded.

Successful treatment and symptom improvement have been shown to lead to normalization of neural functioning.[19,20]

In the event that a study reported contrasts of different UD groups (eg, psychotic and nonpsychotic UD) against the same control group, only the results of the group that was most similar to those investigated in most of the studies (eg, nonpsychotic) were included. Therefore, we avoided inclusion of results that are based on the same control group and possibly reflect peculiarities of these groups. Analyses of group contrasts across several patient groups (ie, main effect of group comparing patients with UD, patients with schizophrenia, and controls) were only included when reporting post hoc results specific to the UD (vs control) group.

Criteria Related to Experimental Design and Contrasts

We only included experiments that used an emotional or a cognitive task and reported group differences or group × condition interactions in task-related brain activity. Consequently, resting-state experiments and those reporting correlations and interactions with other variables (eg, group × performance interaction, correlation with clinical parameters) were excluded because they reflect task-unrelated and strongly specific effects, respectively. To keep the included contrasts and tasks as homogenous as possible we applied the following criteria.

First, emotional tasks were operationally defined as involving presentation of an emotional visual or auditory stimulus. At the contrast level, only group differences (UD vs controls) or group × condition interactions in an emotional vs nonemotional condition were included, whereas experiments investigating group effects (or interactions) between 2 emotional conditions (ie, sad vs happy) were excluded. These latter experiments report specific valence effects that strongly depend on the valence of the target and subtraction condition and cancel out general emotional processes. In addition, this exclusion criterion avoids multiple contrasts of the same participant groups (see below).

Furthermore, we excluded emotional regulation experiments, given that most studies contrasted a reappraisal against an emotional viewing condition and focused predominantly on regulatory mechanisms rather than emotional processing. These tasks can hence be expected to yield brain activations associated with cognitive regulation, whereas emotion-related regions are attenuated.[21] In addition, experiments focusing on anticipatory processing of emotions were excluded owing to the low number of available studies (n = 1).

Second, cognitive tasks were operationally defined as involving a cognitive paradigm. At the contrast level, experiments were included that reported group differences (or interactions) between patients with UD and controls in a cognitive challenge compared with a control (less challenge or baseline) condition. Experiments investigating error-related activity were excluded (n = 3). This criterion was applied because most included cognitive experiments focused on activity in response to predominantly correct responses.

Studies reporting pharmacologic or psychological UD treatment effects were only included if they reported between-group differences at baseline or main effects of diagnosis. Treatment × group interactions were not considered.

To minimize the possibility that meta-analytic results are driven by within-group effects,[22] we limited the contribution of a particular group of participants to 1 experiment per class (with 2 classes for cognitive [increase and decrease] and 4 classes for emotional [increase, decrease, and positive and negative valence] tasks). Hence, if a study reported more than 1 contrast within the same class, these findings were pooled into a single experiment. For example, when a study reported between-group effects in response to angry and fearful faces, the coordinates of the 2 effects were coded as a single negative experiment (information on where this criterion applied is given in eTable 2B in the Supplement).

Criteria Related to Technical Aspects

We included only experiments that reported results of whole-brain group analyses as coordinates in a standard reference space (Talairach[23] or Montreal Neurological Institute [MNI][24]). We excluded studies with results obtained in ROI analyses (n = 17), experiments not covering the whole brain during image acquisition (n = 3), or experiments masking the group differences (or derived from a conjunction) with another contrast (n = 5) (the eMethods in the Supplement provides a detailed explanation).

Because the standard templates used in SPM (Statistical Parametric Mapping) since version SPM96 and FSL (FMRIB Software Library) are in MNI space, coordinates of experiments using SPM or FSL were treated as MNI coordinates if the authors did not explicitly report a transformation from MNI to Talairach space or the use of a different brain template.

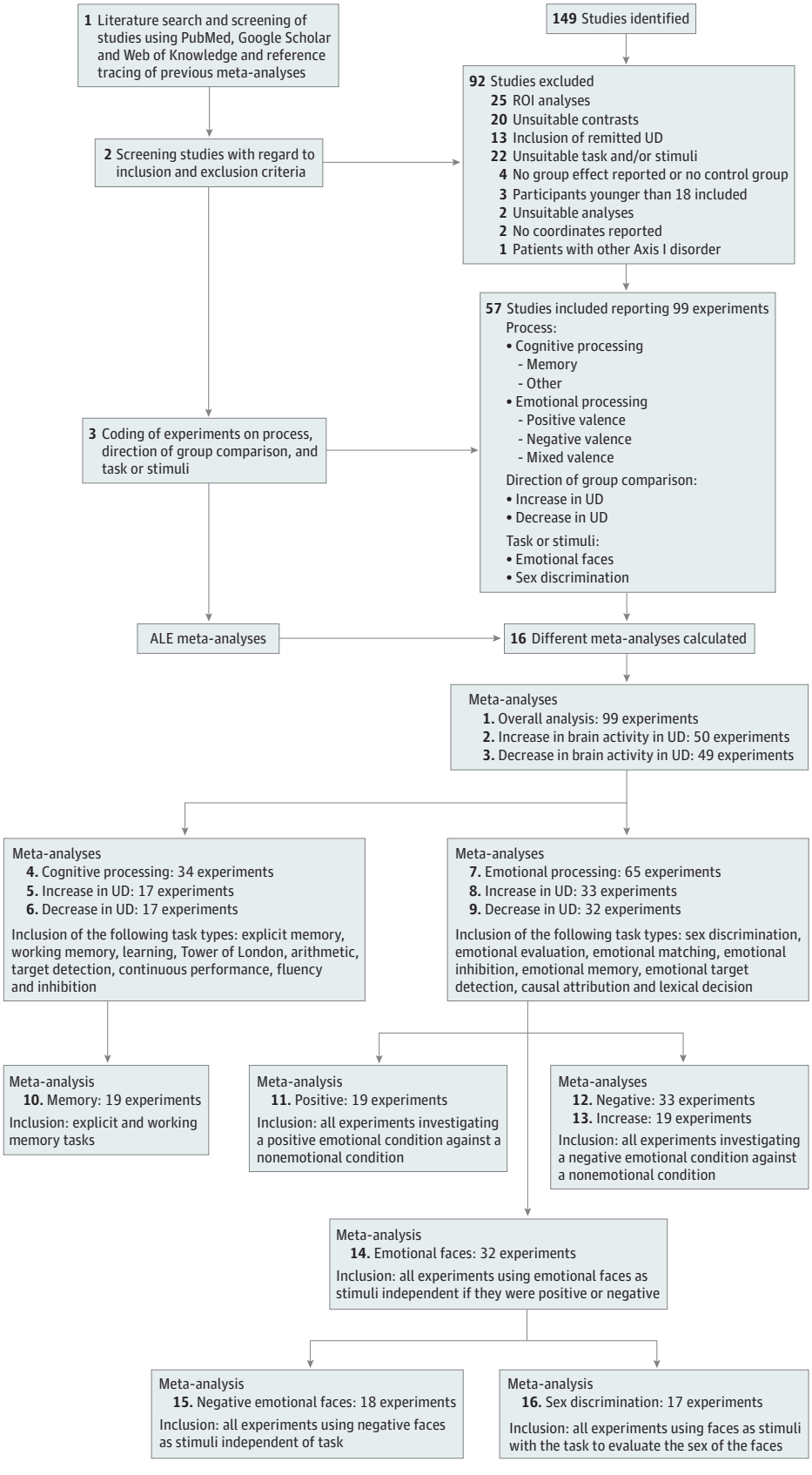
These criteria resulted in inclusion of 57 studies with 99 different experiments (eTable 2 in the Supplement and **Figure 3**).

Different Meta-analytic Groupings

We performed 16 different meta-analyses (Figure 3), with 12 of them corresponding to previously investigated meta-analytic questions (results of the overall analyses are reported in the eResults in the Supplement). We first calculated analyses across experiments independently if they reported increased or decreased brain activity in UD (compared with controls) (see eMethods in the Supplement for an explanation of the advantages of the pooled analysis). Because most previous meta-analyses reported their results separately for increases and decreases, we performed these analyses as well.

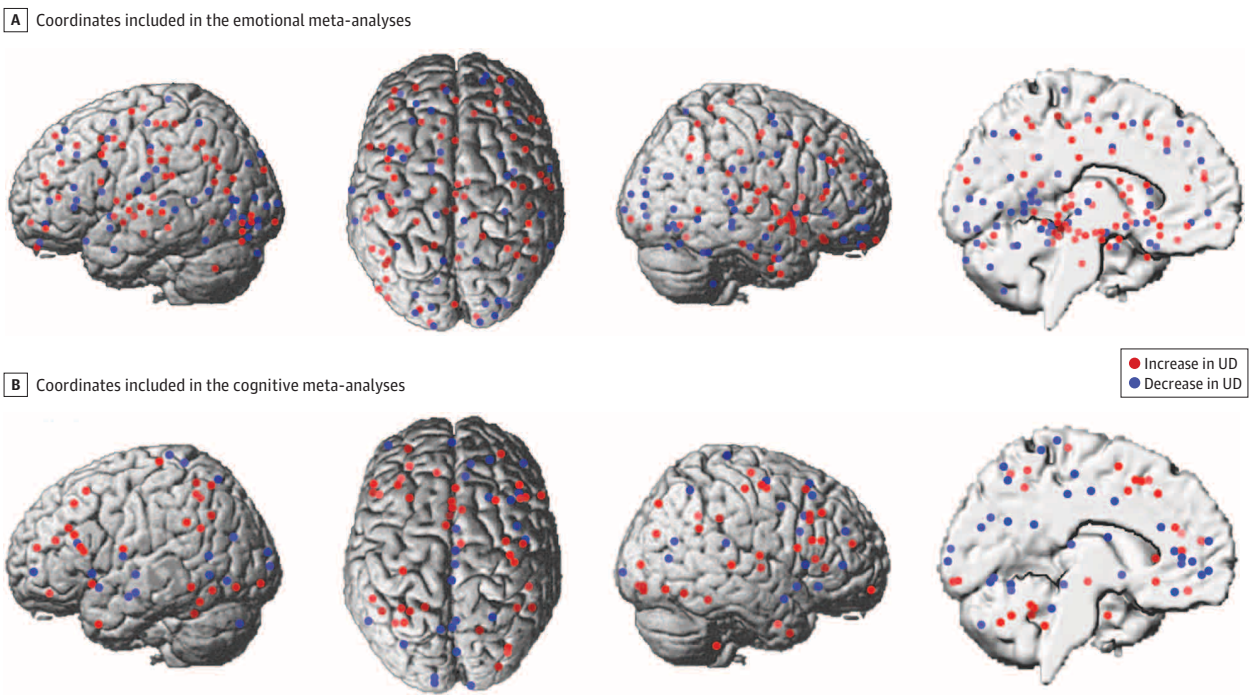
Separate meta-analyses were only performed when a sufficient number of experiments were available (>17 experiments).[16] Thus, analyses differentiating between increases and decreases in UD were not performed for sex discrimination, memory, face stimuli, and positive processing, and for negative processing only a subanalysis for increases was calculated (Figure 3). All analyses (except those with <17 experiments) were repeated to examine (1) patients not receiving medication, (2) patients without comorbidities, (3) patients without late-life or geriatric depression, and (4) corrected results.

Figure 3. Flowchart of the Different Steps Conducted



In total 16 different meta-analyses including 57 studies and 99 experiments were calculated. ALE indicates activation likelihood estimation; ROI, region of interest; and UD, unipolar depression.

Figure 4. Render of the Distribution of Coordinates Included in the Present Meta-analysis



Red indicates increased activation in unipolar depression (UD), whereas blue shows all foci reported as decreased activity in UD. Color transparency indicates the depth of the coordinate related to the cortical surface.

Activation Likelihood Estimation

The ALE meta-analyses[22,25-27] were conducted according to standard procedures used previously.[28] A detailed description can be found in the eMethods in the Supplement and Figure 1. All results were thresholded at a cluster-level corrected threshold of $P < .05$ (cluster-forming threshold at voxel-level $P < .001$).[16]

Results

A total of 57 studies with 99 individual neuroimaging experiments, comprising 1058 patients were included in this analysis. There were 34 cognitive processing experiments and 75 emotional processing experiments; 50 experiments reported increased brain activity in UD, and 49 experiments reported decreased brain activity in UD.

Meta-analyses Across Emotional Experiments

None of the 9 emotional meta-analyses revealed any significant results (all emotional: 65 experiments [$P > .69$]; increases: 33 experiments [$P > .47$]; decreases: 32 experiments [$P > .58$]; negative valence: 33 experiments [$P > .76$]; negative valence increases: 19 experiments [$P > .12$]; positive valence: 19 experiments [$P > .15$]; emotional faces: 32 experiments [$P > .80$]; negative emotional faces: 18 experiments [$P > .75$]; sex discrimination: 17 experiments [$P > .41$]). **Figure 4**A displays the distribution of foci of the emotional analyses.

Meta-analyses Across Cognitive Experiments

None of the 4 cognitive meta-analyses revealed any significant results (all cognitive: 34 experiments [$P > .63$]; increases: 17 experiments [$P > .29$]; decreases: 17 experiments [$P > .97$]; memory: 19 experiments [$P > .48$]). Figure 4B displays the distribution of foci of the cognitive analyses.

Meta-analyses Controlling for Confounds

Analyses restricted to (1) patients not receiving medication, (2) patients without comorbidity, and (3) patients without late-life or geriatric depression revealed similar results (eResults in the Supplement). When restricting the analyses to experiments using corrected statistics (COR), the analyses across experiments of negative emotional processing revealed significant convergence in the left thalamus extending into hippocampus ($x = -18$, $y = -36$, $z = -4$; 5 experiments contributing). All other analyses did not reveal significant convergence (COR all emotional: 38 experiments [$P > .82$]; COR increases emotional: 20 experiments [$P > .27$]; COR decreases emotional: 18 experiments [$P > .23$]; COR all cognitive: 23 experiments [$P > .61$]) (eResults in the Supplement provides details).

Discussion

Inconsistency of Neuroimaging Experiments in UD

Our most important result is the lack of significant convergence in almost all meta-analyses, which should not be attributable to a lack of statistical power. In particularly, Eickhoff

et al[16] recently showed that for clusterwise corrected ALE meta-analyses, a minimum of 17 experiments is needed to achieve power of 80% to detect an effect occurring in one-third of the underlying population of experiments. This criterion was met by all performed analyses, with most being substantially larger and hence holding sufficient power to detect more subtle effects. Moreover, despite the stringent inclusion and exclusion criteria, our work represents, to our knowledge, the largest meta-analysis of task-based neuroimaging experiments in UD to date.

We thus argue that our results indicate a lack of (spatial) convergence among neuroimaging findings in UD. Such heterogeneity and failure to confirm previous effects may be attributable to different factors, including experimental flexibility, that is, differences in experimental design and procedures.[29] For example, emotional processing (65 experiments) was tested by presenting faces (32 experiments [49.2%]), images (14 experiments [21.6%]), words (13 experiments [20%]), and other stimuli (6 experiments [9.2%]); in these experiments, participants were instructed to passively watch or listen (15 experiments [23.1%]) or to evaluate the presented emotion (10 experiments 15.4%]) or sex (17 experiments [26.2%]) or had another task (23 experiments [35.4%]) (eTable 2B in the Supplement provides a detailed description of each experiment). In further acknowledgment of the heterogeneity of investigated emotions and other variations in experimental settings (stimulus duration, intertrial intervals, etc), the lack of convergence may be attributed to experimental differences across experiments. Thus, our results may reflect that the current imaging literature on UD is so heterogeneous that no generalized effects may be found. For example, no general effect in emotional processing in UD may exist, but rather a different impairment may be associated with explicit judgment of the emotion in faces compared with passive observation of emotional faces. The different tasks and stimuli used might thus lead to different effects, which are then not consistent across studies. Unfortunately, too few experiments in UD are similar enough to enable more specific analyses that would definitely address such ambiguity. Thus, concrete conclusions about the neurobiology of aberrant emotional and cognitive processing in UD can only be drawn when enough experiments are similar in procedure and analyzed sample to calculate more specific meta-analyses. In our view, a key contribution to this situation is that new experiments that are too similar to previously published ones are harder to navigate successfully through the peer-review process. Scientists are incentivized to design innovative procedures that differ from previous reports, resulting in a wide range of isolated findings on various paradigms and procedures that lack consistency and do not generalize.

An independent source of heterogeneity is provided by differences in the investigated populations,[30] for example, regarding medication,[31,32] age, prevalence of UD subtypes (eg, chronic vs first-episode), and comorbidities.[33] Specifically, the latter aspect is variable in UD neuroimaging, with 51 (51.5%) of the experiments included in the present analyses excluding all individuals with any other Axis I diagnosis, others excluding specific comorbidities (eg, psychosis), and many not providing sufficient detail on the treatment of individuals with comorbidities. Furthermore, small sample sizes in some experiments[34] might have additionally contributed to inconsistent results.

Finally, analytical flexibility, such as the choices of analysis software, preprocessing parameters, and most important, the specific contrasts calculated (ie, interactions or main effects), further adds to the heterogeneity of the current literature. We should highlight that 38 (38.4%) of all 99 experiments included in our analyses performed statistical inference without correction for multiple comparisons that additionally used various different thresholds. Although Lieberman and Cunningham[35] have argued that uncorrected inference is more sensitive to meaningful (small) effects, it also contaminates the literature with false-positive findings.[36] Unfortunately, taking into account an existing publication bias with negative results being less likely to be published compared with positive ones,[37] such false-positive findings based on invalid inference may actually be more likely to be published than null findings using valid inference. The only significant cluster we found was when we restricted the analyses to just corrected results. This finding additionally highlights the need for functional magnetic resonance imaging studies to correct for multiple comparisons and avoid false-positive findings. In summary, results indicate that neuroimaging results of UD are not consistent across experiments, which is most likely owing to a combination of experimental flexibility, heterogeneity across samples, and the widespread practice of uncorrected inference as well as publishing biases.

Failure to Replicate Meta-analyses

The observed lack of convergence across neuroimaging experiments of UD is at odds with the fact that previous meta-analyses in UD yielded significant findings.[9-14] However, considering the inconsistent findings across these previous meta-analyses, we argue that the current null result might actually reflect the most faithful representation of the current task-based neuroimaging evidence in UD. Nevertheless, the fact that not only individual studies but also meta-analyses in UD have provided inconsistent findings is troubling. Reasons for this inconsistency may be methodological problems of previous analyses or the slightly varying scopes of them, with some focusing on general emotional or cognitive processing[14]; others focusing specifically on negative,[9,11] sad,[13] or positive[9] emotions; and others pooling cognitive and affective processing.[10,12] Although this methodological inconsistency should certainly explain some of the variability, the fact that we computed analyses for almost every scope previously addressed (except sad processing owing to a low number of experiments) but could not replicate any finding indicates a more generalized problem. Thus, based on the present results, we conclude that inconsistencies across previous meta-analyses are not owing to differences in their research question but rather to problems and differences in their methods.

One factor seems to be the trade-off between robustness of the meta-analysis and heterogeneity and quality of the included experiments. Including more experiments ensures that convergence is not driven by single experiments and provides higher power for smaller effects,[16] but often also compromises the homogeneity of the included experiments (eg, pooling different tasks) and their quality (eg, including underpowered experiments). We tried to achieve a balance between

homogeneity and robustness by setting strict inclusion and exclusion criteria. However, these efforts might have contributed to different results compared with previous meta-analyses in UD. For example, Graham and colleagues[12] included resting-state data and Diener and colleagues[10] included experiments in remitted UD, whereas others (including the present study) excluded these. Unfortunately, however, a detailed description of inclusion and exclusion criteria is often lacking in publications of meta-analytic results. A related problem that complicates comparison across meta-analyses is that most do not report in detail which specific experiments and contrasts were included, but only list the included studies. Given that most studies report several contrasts, just listing a study as included does not allow for reproducible meta-analyses.

Another contribution to the inconsistency of previous meta-analyses and their discrepancy to the present null result might be less stringent exclusion criteria regarding experiments that do not reflect whole-brain analyses (eTable 1 in the Supplement). As mentioned, given that the null space in ALE is the entire brain, this aspect is crucial to render the analysis unbiased. In turn, inclusion of such studies will yield a self-fulfilling prophecy because regions that are more often specifically investigated will tend to show artificially high convergence. Thus, some previous meta-analyses might be biased by experiments that do not investigate aberrant activity in UD across the whole brain.

Still the biggest contribution is likely technical in nature. One important aspect is the small sample sizes, particularly in respect to the earliest neuroimaging meta-analyses. In addition to having low power and generalizability, an ALE-specific problem of small sample sizes is that in such analyses convergence is often driven by only 1 or 2 experiments.[16] Thus, previous meta-analyses in UD across a small number of experiments (eg, Fitzgerald and collegues[9] included <7 experiments) might have revealed convergence that is largely attributable to a single experiment. Moreover, we note that most of the previous meta-analyses[9,10,12-14] in UD controlled for the FDR. However, conventional (voxel-wise) FDR correction is not appropriate for inference on neuroimaging analyses such as ALE,[16] because topological inference on spatially smooth data (eg, ALE maps) may lead to spurious clusters.[15] In particular, the combination of low sample size and FDR thresholding may have rendered previous meta-analyses very liberal, leading to an excessive emphasis on apparent convergence across the literature. Furthermore, the FDR correction in GingerALE prior to version 2.3.2 featured a bug,[38] which could have further exaggerated the problem of overly lenient thresholds in the previous meta-analyses on UD.

Recommendations and Outlook

Our results not only indicate inconsistencies across individual experiments investigating aberrant brain activity in UD, but in addition point to problems related to replication of neuroimaging meta-analyses. We suggest that this situation is not specific to UD and functional imaging because in 2011 and 2012, 3 different meta-analyses on structural changes in autism[39-41] were published that demonstrated some correspondence but importantly also some discrepancies. As outlined above, these discrepancies may relate to various conceptual and technical factors, but definite conclusions are difficult given the lack of descriptions at a level of detail that would enable full reproduction.

For clinical neuroimaging we would thus recommend a stronger focus on replication studies, rather than designing complex and newer paradigms. Furthermore, to make replication studies possible, authors must clearly report the specific characteristics of their sample (ie, comorbidity, age range, medication, and chronicity) and correct their results for multiple comparisons.

Neuroimaging meta-analysis is still a maturing yet rapidly developing field. As such, we consider the current case as motivation to formulate basic recommendations for future meta-analyses:

- Meta-analyses should be calculated across a reasonable amount of experiments[16] but concurrently be as homogeneous as possible with respect to the process investigated.
- The inclusion of explicit ROI analyses must be avoided, and attention should also be dedicated to exclude other cases of restricted analysis space.
- Voxel-level FDR thresholding is not a valid approach for statistical inference on smooth data, including neuroimaging meta-analyses.
- Meta-analyses should provide detailed inclusion and exclusion criteria, including a motivation for making these choices.
- Reporting standards must be improved to allow full reproducibility,[42] which includes listing the included studies and specific contrasts.

Last, we note that although our results indicate inconsistencies in results of clinical neuroimaging and meta-analyses, we are still optimistic for the future of neuroimaging. In particular, reproducibility, replication, and data sharing have gained increased attention in the last few years. Our results and recommendations should thus contribute to the awareness of the importance of reproducing previous results.

Conclusions

The present study not only indicates inconsistencies across results of neuroimaging experiments in UD, but also points to replication problems of neuroimaging meta-analyses. These problems highlight the importance of replicating previous results of clinical neuroimaging studies and emphasize the need for better reporting and analysis standards (eg, no inclusion of ROI studies, no FDR correction) for meta-analyses.

ARTICLE INFORMATION

Accepted for Publication: September 7, 2016.

Published Online: November 9, 2016. doi:10.1001/jamapsychiatry.2016.2783

Author Contributions: Dr Müller had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Müller, Cieslik, Serbanescu, Fox, Eickhoff.
Acquisition, analysis, or interpretation of data: Müller, Cieslik, Serbanescu, Laird, Eickhoff.
Drafting of the manuscript: Müller.
Critical revision of the manuscript for important intellectual content: Cieslik, Serbanescu, Laird, Fox, Eickhoff.

Statistical analysis: Müller, Cieslik, Eickhoff.
Obtained funding: Fox, Eickhoff.
Administrative, technical, or material support: Laird, Fox.
Study supervision: Eickhoff.

Conflict of Interest Disclosures: None reported.

Funding/Support: This study was supported by grants EI 816/4-1 and LA 3071/3-1 from the Deutsche Forschungsgemeinschaft, grant R01-MH074457 from the National Institute of Mental Health, the Helmholtz Portfolio Theme "Supercomputing and Modelling for the Human Brain," and grant agreement 604102 from the European Union Seventh Framework Program FP7/2007-2013.

Role of Funder/Sponsor: The funding sources had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015;386(9995):743-800.

2. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)*. Geneva, Switzerland: World Health Organization; 2010.

3. Snyder HR. Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. *Psychol Bull*. 2013;139(1):81-132.

4. Kohler CG, Hoffman LJ, Eastman LB, Healey K, Moberg PJ. Facial emotion perception in depression and bipolar disorder: a quantitative review. *Psychiatry Res*. 2011;188(3):303-309.

5. Fitzgerald PB, Srithiran A, Benitez J, et al. An fMRI study of prefrontal brain activation during multiple tasks in patients with major depressive disorder. *Hum Brain Mapp*. 2008;29(4):490-501.

6. Canli T, Sivers H, Thomason ME, Whitfield-Gabrieli S, Gabrieli JD, Gotlib IH. Brain activation to emotional words in depressed vs healthy subjects. *Neuroreport*. 2004;15(17):2585-2588.

7. Müller VI, Cieslik EC, Kellermann TS, Eickhoff SB. Crossmodal emotional integration in major depression. *Soc Cogn Affect Neurosci*. 2014;9(6):839-848.

8. 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.

9. Fitzgerald PB, Laird AR, Maller J, Daskalakis ZJ. A meta-analytic study of changes in brain activation in depression. *Hum Brain Mapp*. 2008;29(6):683-695.

10. Diener C, Kuehner C, Brusniak W, Ubl B, Wessa M, Flor H. A meta-analysis of neurofunctional imaging studies of emotion and cognition in major depression. *Neuroimage*. 2012;61(3):677-685.

11. Hamilton JP, Etkin A, Furman DJ, Lemus MG, Johnson RF, Gotlib IH. Functional neuroimaging of major depressive disorder: a meta-analysis and new integration of base line activation and neural response data. *Am J Psychiatry*. 2012;169(7):693-703.

12. Graham J, Salimi-Khorshidi G, Hagan C, et al. Meta-analytic evidence for neuroimaging models of depression: state or trait? *J Affect Disord*. 2013;151(2):423-431.

13. Lai CH. Patterns of cortico-limbic activations during visual processing of sad faces in depression patients: a coordinate-based meta-analysis. *J Neuropsychiatry Clin Neurosci*. 2014;26(1):34-43.

14. Palmer SM, Crewther SG, Carey LM; START Project Team. A meta-analysis of changes in brain activity in clinical depression. *Front Hum Neurosci*. 2015;8:1045.

15. Chumbley JR, Friston KJ. False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage*. 2009;44(1):62-70.

16. Eickhoff SB, Nichols TE, Laird AR, et al. Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *Neuroimage*. 2016;137:70-85.

17. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed, text revision. Washington, DC: Amercian Psychiatric Association; 2000.

18. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Washington, DC: American Psychiatric Association; 2013.

19. Dichter GS, Felder JN, Smoski MJ. The effects of brief behavioral activation therapy for depression on cognitive control in affective contexts: an fMRI investigation. *J Affect Disord*. 2010;126(1-2):236-244.

20. Ritchey M, Dolcos F, Eddington KM, Strauman TJ, Cabeza R. Neural correlates of emotional processing in depression: changes with cognitive behavioral therapy and predictors of treatment response. *J Psychiatr Res*. 2011;45(5):577-587.

21. Buhle JT, Silvers JA, Wager TD, et al. Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cereb Cortex*. 2014;24(11):2981-2990.

22. Turkeltaub PE, Eickhoff SB, Laird AR, Fox M, Wiener M, Fox P. Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Hum Brain Mapp*. 2012;33(1):1-13.

23. Talairach J, Tournoux P. *Co-planar Stereotaxic Atlas of the Human Brain*. New York, NY: Thieme; 1988.

24. Collins DL, Neelin P, Peters TM, Evans AC. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J Comput Assist Tomogr*. 1994;18(2):192-205.

25. Eickhoff SB, Laird AR, Grefkes C, Wang LE, Zilles K, Fox PT. Coordinate-based ALE meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum Brain Mapp*. 2009;30(9):2907-2926.

26. Eickhoff SB, Bzdok D, Laird AR, Kurth F, Fox PT. Activation likelihood estimation meta-analysis revisited. *Neuroimage*. 2012;59(3):2349-2361.

27. Turkeltaub PE, Eden GF, Jones KM, Zeffiro TA. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage*. 2002;16(3 Pt 1):765-780.

28. Cieslik EC, Mueller VI, Eickhoff CR, Langner R, Eickhoff SB. Three key regions for supervisory attentional control: evidence from neuroimaging meta-analyses. *Neurosci Biobehav Rev*. 2015;48:22-34.

29. Maxwell SE, Lau MY, Howard GS. Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *Am Psychol*. 2015;70(6):487-498.

30. Doose-Grünefeld S, Eickhoff SB, Müller VI. Audiovisual emotional processing and neurocognitive functioning in patients with depression. *Front Integr Neurosci*. 2015;9:3.

31. Davidson RJ, Irwin W, Anderle MJ, Kalin NH. The neural substrates of affective processing in depressed patients treated with venlafaxine. *Am J Psychiatry*. 2003;160(1):64-75.

32. Fu CH, Williams SC, Cleare AJ, et al. 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*. 2004;61(9):877-889.

33. Cusi AM, Nazarov A, Holshausen K, Macqueen GM, McKinnon MC. Systematic review of the neural basis of social cognition in patients with mood disorders. *J Psychiatry Neurosci*. 2012;37(3):154-169.

34. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365-376.

35. Lieberman MD, Cunningham WA. Type I and type II error concerns in fMRI research: re-balancing the scale. *Soc Cogn Affect Neurosci*. 2009;4(4):423-428.

36. Bennett CM, Wolford GL, Miller MB. The principled control of false positives in neuroimaging. *Soc Cogn Affect Neurosci*. 2009;4(4):417-422.

37. Ioannidis JP, Munafò MR, Fusar-Poli P, Nosek BA, David SP. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn Sci*. 2014;18(5):235-241.

38. Eickhoff SB, Laird AR, Fox PM, Lancaster JL, Fox PT. Implementation errors in the GingerALE Software: description and recommendations [published online August 11, 2016]. *Hum Brain Mapp*. doi:10.1002/hbm.23342

39. Duerden EG, Mak-Fan KM, Taylor MJ, Roberts SW. Regional differences in grey and white matter in children and adults with autism spectrum disorders: an activation likelihood estimate (ALE) meta-analysis. *Autism Res*. 2012;5(1):49-66.

40. Nickl-Jockschat T, Habel U, Michel TM, et al. Brain structure anomalies in autism spectrum disorder: a meta-analysis of VBM studies using anatomic likelihood estimation. *Hum Brain Mapp*. 2012;33(6):1470-1489.

41. Via E, Radua J, Cardoner N, Happé F, Mataix-Cols D. Meta-analysis of gray matter abnormalities in autism spectrum disorder: should Asperger disorder be subsumed under a broader umbrella of autistic spectrum disorder? *Arch Gen Psychiatry*. 2011;68(4):409-418.

42. Nichols TE, Das S, Eickhoff SB, et al. Best practices in data analysis and sharing in neuroimaging using MRI. http://biorxiv.org/content/early/2016/05/20/054262. Posted May 20, 2016. Accessed May 28, 2016.