ONLINE FIRST

Excess Significance Bias in the Literature on Brain Volume Abnormalities

John P. A. Ioannidis, MD, DSc

Context: Many studies report volume abnormalities in diverse brain structures in patients with various mental health conditions.

Objective: To evaluate whether there is evidence for an excess number of statistically significant results in studies of brain volume abnormalities that suggest the presence of bias in the literature.

Data Sources: PubMed (articles published from January 2006 to December 2009).

Study Selection: Recent meta-analyses of brain volume abnormalities in participants with various mental health conditions vs control participants with 6 or more data sets included, excluding voxel-based morphometry.

Data Extraction: Standardized effect sizes were extracted in each data set, and it was noted whether the results were "positive" ($P<.05$) or not. For each data set in each meta-analysis, I estimated the power to detect at $\alpha=.05$ an effect equal to the summary effect of the respective meta-analysis. The sum of the power estimates gives the number of expected positive data sets. The expected number of positive data sets can then be compared against the observed number.

Data Synthesis: From 8 articles, 41 meta-analyses with 461 data sets were evaluated (median, 10 data sets per meta-analysis) pertaining to 7 conditions. Twenty-one of the 41 meta-analyses had found statistically significant associations, and 142 of 461 (31%) data sets had positive results. Even if the summary effect sizes of the meta-analyses were unbiased, the expected number of positive results would have been only 78.5 compared with the observed number of 142 ($P<.001$).

Conclusion: There are too many studies with statistically significant results in the literature on brain volume abnormalities. This pattern suggests strong biases in the literature, with selective outcome reporting and selective analyses reporting being possible explanations.

Arch Gen Psychiatry. 2011;68(8):773-780.
Published online April 4, 2011.
doi:10.1001/archgenpsychiatry.2011.28

Author Affiliations: Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, and Biomedical Research Institute, Foundation for Research and Technology–Hellas, Ioannina, Greece; Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, and Department of Medicine, Tufts University School of Medicine, and Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts; and Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, California.

BRAIN VOLUME ABNORMALITIES have been associated with a large variety of mental health diseases and conditions[1-12] and have typically been a key topic in the discussion of the pathophysiology of mental disorders for the past 25 years.[13-15] The literature on brain volume abnormalities is rapidly expanding, with hundreds of studies published to date. A considerable number of meta-analyses have already been published that try to summarize the results from these studies of brain volume abnormalities.[1-12] These meta-analyses identify significant associations for specific brain volumes and structures for almost any disease and condition assessed, including schizophrenia, major depression, bipolar disorder, posttraumatic stress disorder, obsessive-compulsive disorder, autism, and personality disorders.[1-8]

The large number of statistically significant associations could have several explanations. One possibility is that all major mental conditions have genuine correlates with brain volumes. Some associations may indicate specific conditions, whereas others may be seen in very diverse diseases. Another possibility is that reporting bias is operating in the literature.[16,17] Reporting bias could include the following mechanisms: (1) study publication bias, in which the results of nonstatistically significant ("negative") studies are left unpublished; (2) selective outcome reporting bias, in which results of outcomes (in this case, the volume of specific brain structures) that are negative are left unpublished, whereas the "positive" associations with other brain volumes are published; and (3) selective analysis reporting bias, in which data on the volume of a particular brain structure are analyzed with different methods and in which positive results are preferentially published over negative results. The common denominator of all these mechanisms is that the published literature has an excess of statistically significant results (ie, excess significance bias).[18]

Detecting these biases is not a straightforward process. Many meta-analyses in this field have applied asymmetry (funnel plot) tests that determine whether small studies give different results from larger studies.[19] If so, such small-study effects may be due to publication bias or other reporting bias. However, these tests are neither sensitive nor specific for detecting reporting biases.[20] In the literature on brain volume abnormalities, they may be particularly unsuitable because all studies have limited sample sizes, thus the range of "large" vs "small" studies is too narrow.[21] Moreover, there are few studies that report data on a particular brain structure, and asymmetry tests do not work well when there are fewer than 10 to 20 studies.[20,21]

A more appropriate alternative is to apply an excess significance test that specifically evaluates whether there are too many reported studies that have statistically significant results.[18] This test has the additional advantage that it can evaluate the excess of significant studies not only in a single meta-analysis but across many meta-analyses in a given field. This could include all meta-analyses of brain volumes for a given condition or all meta-analyses of brain volumes for many different conditions. Herein, I have applied such a test to evaluate whether the literature on brain volume abnormalities is subject to excess significance bias.

METHODS

EVALUATED DATA

Data were collected from recent comprehensive meta-analyses of studies comparing participants with specific mental health conditions vs control participants for differences in brain volumes of specific brain structures. I focused on volumetric studies and did not consider meta-analyses that used measures of gray matter density derived from voxel-based morphometry. After perusing PubMed, I selected articles using the following search strategy: brain volume AND meta-analysis. I limited the search to human studies (last search December 20, 2009) and focused only on meta-analyses published from 2006 to 2009, because earlier articles would be likely to miss a large number of recent studies. The full text of potentially eligible articles was scrutinized. Of those, articles were retained if they included at least 1 meta-analysis for volume differences in a brain structure in which information was provided or could be calculated per study on the number of participants in each of the 2 compared groups (those with the condition of interest and controls) and the standardized effect size (expressed as Cohen's *d*, Hedges' *g*, or other similar standardized metrics) for the comparison. When more than 1 article on the same condition was eligible and contained usable data, complementary information from more than 1 article was used, if the usable data pertained to volumes of different brain structures; conversely, only the most recent article was retained, if the usable data pertained to the volumes of the same brain structures.

Herein, each study data set corresponds to a separate estimate of effect. The evaluation did not consider total brain and total intracranial volumes because these are very nonspecific measures and are often treated as covariates of no or limited interest. It also excluded meta-analyses with fewer than 6 study data sets; the cutoff decision was made a priori because for most meta-analyses with fewer study data sets, it would have been unlikely to make solid conclusions about the presence or absence of excess significance, given the limited evidence. Meta-analyses were accepted regardless of whether they analyzed separately left/right structures or just the total volume of both sides. When meta-analyses on both the total volume and left/right volumes were presented, either the total or the left/right side meta-analyses were kept, depending on which had a larger sample size; when the sample size was the same, the separate left/right data were preferred. Data were eligible regardless of what imaging technique and technical parameters thereof had been used.

EXCESS SIGNIFICANCE TEST

I used a previously developed test for excess significance.[18,22] In brief, the test evaluates whether the number of single-study data sets that report nominally statistically significant results ($P < .05$) among those included in a meta-analysis is too large based on the power that these data sets have to detect plausible effects at $\alpha = .05$. In each meta-analysis, I calculated the power of each study data set. The sum of the power estimates gives the expected number of positive study data sets (those with nominally statistically significant results). As previously presented in detail,[18,22] the observed number, O, of positive study data sets in each meta-analysis can be compared against the expected number, E, of positive study data sets with a χ^2 test or with a binomial test, and the results are practically equivalent. The O vs E comparison is extended to many meta-analyses, by summing the O and E numbers from each meta-analysis. If there is no excess significance bias, then O = E. The greater the difference between O and E, the greater is the extent of excess significance bias.

The estimated power of each study data set depends on what is the plausible effect size.[18] The true effect size for any meta-analysis cannot be known. In the absence of bias, one would expect the observed (estimated) summary effect size to be a good representation of the true effect sizes, allowing simply for estimation or random error. In the presence of bias, one would expect the observed effect size to be larger than the true effect size, and the divergence would be expected to become larger with an increased level of bias. Thus, one has to consider a range of values for the plausible true effect size that may not be the same as the observed one. Herein, I considered an optimistic scenario, in which the true effect is assumed to be equal to the observed effect, and a more pessimistic scenario, in which the true effect is assumed to be equal to half the observed effect.[23]

CALCULATIONS AND SOFTWARE

All effect estimates were expressed as standardized mean differences for the 2 compared groups, with the metrics chosen by the authors of each original meta-analysis. Effect size computation in each study in each meta-analysis takes into account the mean volume and the variance of the volume in cases and controls, and variances can be different in cases and controls. Summary effects in meta-analyses are based on random-effects calculations. When the standardized mean difference and variance thereof were not given, they were calculated from the provided sample size *n*, mean values *m*, and standard deviations SD of the absolute measurements per each group (1 and 2) using the following formulas[24]:

$$\text{Standardized effect} = \frac{(m_1 - m_2)\left(1 - \frac{3}{4(n_1+n_2)-9}\right)}{\sqrt{\frac{(n_1-1)\text{SD}_1^2 + (n_2-1)\text{SD}_2^2}{n_1+n_2-2}}}.$$

$$\text{Variance} = \frac{n_1+n_2}{n_1 n_2} + \frac{g^2}{2(n_1+n_2-3.94)}.$$

For each meta-analysis, the *P* value of the χ^2-based *Q* test and the I^2 metric of inconsistency were recorded. The Cochran *Q* test[25]

is obtained by the weighted sum of the squared differences of the observed effect in each study minus the fixed summary effect. The I^2 metric[26] is the ratio of the between-study variance over the sum of the within- and between-study variance. When the I^2 was not given, it was calculated from the formula $I^2 = (Q - k + 1)/Q$, where k is the number of studies.[27] The Q test is considered significant here at $P < .05$, but it should be interpreted with caution owing to the possibility of both false positives and false negatives.[28,29] For I^2, as a rough guide, values exceeding 50% are considered as large heterogeneity beyond chance, and values exceeding 75% are considered very large heterogeneity.[27] However, with a limited number of study data sets, the uncertainty in the estimates of I^2 can be substantial, and thus these inferences should be made with great caution.[30]

The PS: Power and Sample Size Calculation program[31] was used to estimate the power of each study. Excess significance is claimed at $P < .05$, and results are also presented with Bonferroni correction for the number of examined conditions and brain structures.

RESULTS

EVALUATED DATA

The search yielded 41 items, of which 22 were excluded after perusing the title and abstract. Of the remaining 19 articles, 3 were voxel-based meta-analyses, and 5 did not provide sufficient details of results per data set. Of the 11 articles with detailed results per data set, 3 addressed conditions and brain areas that had been examined in more recent meta-analyses and were thus excluded to avoid data overlap. Therefore, data from 8 articles were eligible for the analysis, including data on meta-analyses of brain volume abnormalities in major depressive disorder,[1,2] bipolar disorder,[3] obsessive-compulsive disorder,[4] posttraumatic stress disorder,[5] autism,[6] first episode of schizophrenia,[7] and relatives of patients with schizophrenia.[8]

Table 1 summarizes the evaluated data from the 8 eligible meta-analysis publications.[1-8] All meta-analyses include only magnetic resonance imaging studies, except for that of Kempton et al,[3] who also allowed the inclusion of computed tomographic (CT) scan studies. A total of 461 data sets had been included in 41 meta-analyses on brain volumes for 7 different conditions. All studies included in the relevant meta-analyses had been published in peer-reviewed journals, except for 1 study (2 data sets included in the calculations) that had been published as an abstract. There were 6 to 31 data sets per meta-analysis (median, 10 data sets). These were typically small data sets, and cumulatively no meta-analysis had a sample size (cases and controls combined) exceeding 1000, with the exception of the meta-analyses of hippocampus volume in major depressive disorder and in relatives of patients with schizophrenia. In 14 meta-analyses, there were nominally statistically significant differences in larger brain volumes among cases; in 7 meta-analyses, there were nominally statistically significant differences in larger brain volumes among controls; and in 20 meta-analyses, there were no significant differences between the 2 groups. Only 5 effects sizes had an absolute magnitude exceeding 0.50 (anterior cingulated cortex in major depressive disorder as well as left hippocampus, right hippocampus, left lateral ventricle, and third ventricle in first-episode schizophrenia). Of the 41 effects, none have large point estimates (>0.8 in absolute magnitude), and the 95% confidence intervals exclude large effects for 40 of 41 meta-analyses. The 95% confidence intervals also exclude moderate effects (>0.5 in absolute magnitude) in 27 of the 41 meta-analyses. There was nominally statistically significant heterogeneity in 24 of 41 meta-analyses. I^2 values exceeding 50% were noted in 19 of the 41 meta-analyses, and 5 of those had values exceeding 75%.

OBSERVED VS EXPECTED NUMBER OF POSITIVE STUDY DATA SETS

Table 1 also shows the observed and expected number of positive study data sets in each meta-analysis, assuming the plausible effect to be the summary effect of the meta-analysis or half of this effect. For most meta-analyses (29/41), the observed is larger than the expected, and in only 10 meta-analyses is the observed smaller than the expected (2 meta-analyses had an equal observed and expected number of positive study data sets), even if we assume that the summary effect seen in the meta-analysis is the most plausible estimate of the true effect. If the plausible effect is assumed to be half of what is seen in each meta-analysis, then O is larger than E for 36 meta-analyses, whereas the opposite occurs only in 5 meta-analyses with few studies, all of which have O=0 and E=0.3 to 0.6.

If we assume that the summary effect seen in the meta-analysis is the most plausible estimate of the true effect, then 16 of the 41 meta-analyses show evidence ($P < .05$) for an excess O over E and 7 of them show evidence for an excess O over E, even after correcting for the number of tested conditions and brain structures. If the plausible effect is assumed to be half of what is seen in each meta-analysis, then the respective numbers of meta-analyses are 27 and 14.

Meta-analyses that have larger estimates of heterogeneity (as expressed by I^2) tend to also have large differences between O and E (Spearman correlation coefficient $\rho = 0.63$ when plausible effects are considered to be those observed in the summary of the meta-analyses [**Figure**] and $\rho = 0.53$ when plausible effects are considered to be half of those observed).

Table 2 shows the composite data from all meta-analyses for each mental health condition. The observed number of positive study data sets is always larger than the expected, regardless of the assumptions of what the plausible effect should be in each meta-analysis. The difference between E and O is beyond chance ($P < .05$) for 5 of the 7 conditions when the plausible effect is assumed to be the same as the summary effect in each meta-analysis (all except first-episode schizophrenia and relatives of patients with schizophrenia) and for all 7 conditions when the plausible effect is assumed to be half of that magnitude. With Bonferroni correction, the difference between E and O is statistically significant for 4 and 5 of the 7 conditions, respectively.

Table 2 also groups meta-analyses per brain structure. Of the 15 brain structures evaluated, 8 showed a

Table 1. Evaluated Meta-Analyses of Brain Volume Abnormalities and Observed and Expected Number of "Positive" Data Sets[a]

Article, Condition, and Brain Structure	Study Data Sets, No.	Cases/ Controls, No.	Effect Size (95% CI)	*P* Value[b]	I^2, %	Observed Positive Data Sets, No.	Expected Positive Data Sets,[c] No.	Expected Positive Data Sets Under Half-Effect Assumption,[d] No.
Koolschijn et al[1]								
Major depressive disorder								
Anterior cingulate cortex	8	181/170	–0.77 (–1.32 to –0.22)	<.001	84	4	4.9	1.7[e]
Orbitofrontal cortex	7	373/204	–0.43 (–0.78 to –0.09)	.001	73	4	2.9	1.2[e]
Prefrontal cortex	7	242/181	–0.34 (–0.52 to –0.16)	.57	0	1	1.7	0.9
Hippocampus	31	1114/991	–0.41 (–0.54 to –0.28)	<.001	51	14	10.2	3.7[f]
Putamen	6	192/184	–0.48 (–0.80 to –0.16)	.04	58	3	2.6	0.9[e]
Caudate nucleus	10	467/316	–0.31 (–0.58 to –0.04)	.001	70	2	2.4	1.0
Hajek et al[2]								
Major depressive disorder								
Left amygdala	20	409/482	0.04 (–0.21 to 0.28)	<.001	66	6	1.0[f]	1.0[f]
Right amygdala	20	409/482	–0.08 (–0.37 to 0.21)	<.001	76	8	1.1[f]	1.0[f]
Kempton et al[3]								
Bipolar disorder								
Lateral ventricles	17	375/589	0.39 (0.24-0.55)	.24	19	5	4.4	1.8[e]
Third ventricle	12	208/271	0.27 (0.00-0.53)	.04	46	2	1.5	0.8
Gray matter	14	257/310	–0.18 (–0.50 to 0.13)	<.001	69	4	1.2[e]	0.8[f]
White matter	14	221/284	–0.09 (–0.32 to 0.15)	.05	41	1	0.8	0.7
Left caudate nucleus	11	273/273	–0.03 (–0.21 to 0.15)	.36	9	0	0.6	0.6
Right caudate nucleus	11	273/273	–0.07 (–0.24 to 0.10)	.50	0	0	0.6	0.6
Left putamen	7	197/183	–0.02 (–0.22 to 0.18)	.57	0	0	0.4	0.4
Right putamen	7	197/183	0.00 (–0.20 to 0.21)	.63	0	0	0.4	0.4
Globus pallidus	6	135/106	0.50 (0.00-1.01)	<.001	71	2	0.5[e]	0.4[e]
Thalamus	10	235/207	–0.02 (–0.32 to 0.28)	.01	59	3	0.5[f]	0.5[f]
Left temporal lobe	12	258/277	–0.08 (–0.35 to 0.20)	.01	56	3	0.6[e]	0.6[e]
Right temporal lobe	12	258/277	–0.16 (–0.44 to 0.12)	.01	55	3	1.0[e]	0.6[e]
Left hippocampus	18	380/487	0.10 (–0.06 to 0.26)	.23	18	1	1.0	0.9
Right hippocampus	18	380/487	0.02 (–0.13 to 0.17)	.32	11	1	0.9	0.9
Left amygdala	11	236/354	–0.07 (–0.47 to 0.33)	<.001	80	5	0.6[f]	0.6[f]
Right amygdala	11	236/354	–0.04 (–0.45 to 0.37)	<.001	81	6	0.6[f]	0.6[f]
Rotge et al[4]								
Obsessive-compulsive disorder								
Left caudate nucleus	8	159/160	–0.10 (–0.37 to 0.17)	.20	29	1	0.4	0.4
Right caudate nucleus	8	159/160	–0.08 (–0.40 to 0.25)	.05	50	2	0.4[e]	0.4[e]
Karl et al[5]								
Posttraumatic stress disorder								
Right hippocampus	15	250/312	–0.28 (–0.42 to –0.13)	<.001	63	8	1.9[f]	1.0[f]
Left hippocampus	15	250/312	–0.29 (–0.43 to –0.14)	<.001	65	6	2.0[e]	1.0[e]
Right amygdala	7	131/188	–0.07 (–0.21 to 0.07)	.20	27	1	0.4	0.4
Left amygdala	7	131/188	–0.14 (–0.26 to –0.00)	.30	17	2	0.5[e]	0.4[e]
Stanfield et al[6]								
Autism								
Left amygdala	6	109/100	0.15 (–0.46 to 0.76)	<.001	76	4	0.5[f]	0.3[f]
Vermal lobules I-IV	10	290/310	0.10 (–0.28 to 0.49)	<.001	72	3	0.7[e]	0.5[f]
Vermal lobules VI-VII	12	348/337	–0.27 (–0.51 to –0.03)	.02	52	5	1.9[e]	0.9[f]
Steen et al[7]								
First-episode schizophrenia								
Left hippocampus	11	300/287	–0.53 (–0.74 to –0.33)	.18	28	5	4.8	1.6[e]
Right hippocampus	11	300/287	–0.53 (–0.76 to –0.31)	.09	38	5	4.8	1.6[e]
Left lateral ventricle	9	262/248	0.60 (0.42-0.78)	.42	2	6	5.2	1.7[f]
Right lateral ventricle	9	262/248	0.46 (0.28-0.64)	.58	0	4	3.4	1.0[e]
Third ventricle	8	204/209	0.59 (0.39-0.79)	.95	0	6	4.1	1.3[f]
Boos et al[8]								
Relative of patient with schizophrenia								
Hippocampus	9	421/603	–0.31 (–0.49 to –0.13)	.09	42	4	2.9	1.2[e]
Gray matter	7	249/285	–0.18 (–0.33 to –0.02)	.70	0	1	0.8	0.4
Third ventricle	7	414/418	0.21 (0.03-0.40)	.22	28	1	1.2	0.4

Abbreviation: CI, confidence interval.

[a] Those with $P < .05$.

[b] Of the χ^2-based Q test.

[c] The expected number of positive data sets (those with $P < .05$) is the sum of the power of all the studies to detect the assumed plausible effect in each meta-analysis at $\alpha = .05$.

[d] Based on the assumption that the plausible effect is half of what is seen in each meta-analysis.

[e] Nominally statistically significant difference between expected and observed.

[f] Statistically significant difference between expected and observed even after Bonferroni adjustment for the total number of tested conditions and brain structures.

difference between E and O beyond chance ($P<.05$), when the plausible effect is assumed to be the same as the summary effect in each meta-analysis, and this increases to 12 when the plausible effect is assumed to be half of that magnitude. With Bonferroni correction, the difference between E and O remains statistically significant for 5 and 9 of the 15 brain structures, respectively. The larger fold deviation of O from E was seen for amygdala under either assumption.

When data are combined from all 41 meta-analyses, there are 142 observed data sets with nominally statistically significant results among the 461 (31%), whereas the expected number would be 78.5 and 37.1 under the 2 effect assumptions, respectively ($P<.001$ for comparison with the observed for both analyses).

COMMENT

This evaluation of 461 data sets in 41 meta-analyses of brain volumes in diverse conditions shows that, in the literature, the number of positive results is way too large to be true. Even if the effect sizes observed in the meta-analyses are accurate, the number of positive results (n=142) is almost double than what would have been expected based on power calculations for the included samples. If the true effect sizes are only half of those observed in the meta-analyses, then the number of positive results is about 4 times the expected number thereof. Bias may be present in meta-analyses of all 7 examined conditions and in most of the examined brain structures. Such bias threatens the validity of the overall literature on brain volume abnormalities.

The excess significance may be due to unpublished negative results, or it may be due to negative results having been turned into positive results through selective exploratory analyses. If all the excess significance is due to negative results not being published, then this means that only slightly more than 1 in 2 or 1 in 4 negative results have been published, depending on what plausible effect size is assumed. This would correspond to approximately 600 to 1200 unpublished negative results, besides the 319 that have been published. Conversely, the excess significance may be due to negative results becoming positive: given that the expected positive results are 79 or 37, with the 2 analyses, then one can estimate that a negative-to-positive conversion of 64 or 105 results, respectively, among the 142 observed positive ones would suffice to cause this excess of significance. Possibly both mechanisms contribute.

First, bias against the publication of negative results, the traditional form of publication bias,[17] may exist. Some of the prior meta-analyses tried to investigate small-study effects (whether small studies give more prominent results than larger studies), which may signal publication bias. However, this association is nonspecific,[20] and most studies on brain volumes are small anyhow, so differentiating between small and large makes little sense. Moreover, the typical investigation of brain volumes is likely to measure by default the volume of multiple brain structures. Because of multiple comparisons, most investigations may have at least 1 positive result to report,

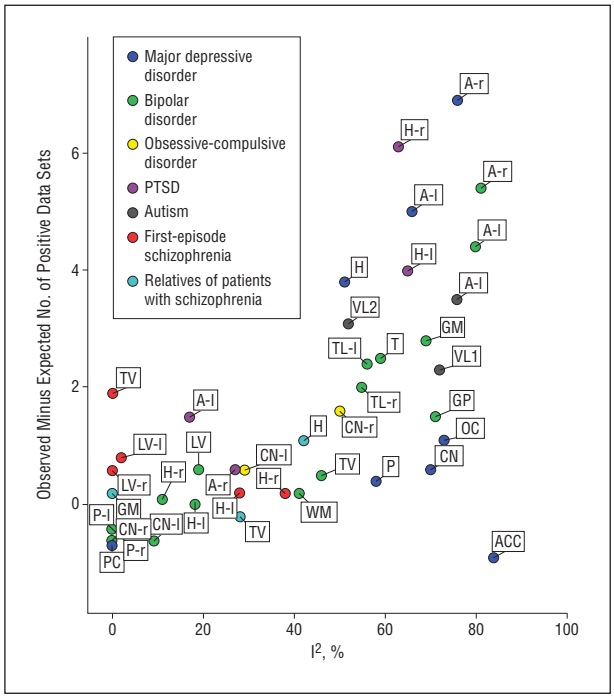

Figure. Correlation between observed minus expected number of "positive" study data sets with the between-study heterogeneity (I^2). Different colors are shown for each condition, and brain structures are also labeled on the plot. A indicates amygdala; ACC, anterior cingulated cortex; CN, caudate nucleus; GM, gray matter; GP, globus pallidus; H, hippocampus; l, left; LV, lateral ventricles; OC, orbitofrontal cortex; P, putamen; PF, prefrontal cortex; PTSD, posttraumatic stress disorder; r, right; T, thalamus; TL, temporal lobe; TV, third ventricle; VL1, vermal lobules I-IV; VL2, vermal lobules VI-VII; and WM, white matter.

even if this is only a chance finding due to an inflated type I error. This suggests that bias is more likely to occur at the level of outcome reporting (ie, with only a subset of the brain regions, among the many evaluated, being reported in the published article, rather than the whole study remaining unpublished). The most suggestive evidence for this type of outcome reporting bias in the literature comes from the mere juxtaposition of the availability of information for different brain regions in studies addressing the same condition. Some brain regions have data reported from far more studies than others. For example, although there are 31 reported results on hippocampus volume in studies of major depression, only 6 of them report on putamen volumes, and only 7 report on orbitofrontal cortex or prefrontal cortex volumes.[1] To some extent, this difference may reflect the fact that investigators genuinely focused only on the hippocampus in some studies or that interest in hippocampus abnormalities preceded interest in the study of other volumes. However, it is possible that many studies did measure comprehensively all or many of the major brain areas and reported selectively on a few, with reporting guided in part by the significance of the results.

Second, one suspects that some analyses that were negative have been presented as positive. This type of bias has been best documented in randomized trials (eg, trials of antidepressants).[31] As noted earlier, selective analysis reporting bias can have a very influential effect on the results: one must convert relatively few studies from nega-

Table 2. Observed and Expected Number of "Positive" Study Data Sets Across All Meta-analyses for Each Condition and for Each Brain Structure

	Study Data Sets, No.	Observed Positive Data Sets, No.	Expected Positive Data Sets,[a] No.	Expected Postive Data Sets Under Half-Effect Assumption,[b] No.
According to condition				
Major depressive disorder	109	42	26.8[c]	11.4[d]
Bipolar disorder	191	36	15.6[c]	11.2[c]
Obsessive-compulsive disorder	16	3	0.8[d]	0.8[d]
Posttraumatic stress disorder	44	17	4.8[c]	2.8[c]
Autism	28	12	3.1[c]	1.7[c]
First-episode schizophrenia	48	26	22.3	7.2[c]
Relative of patient with schizophrenia	25	6	4.9	2.0[d]
According to brain structure				
Anterior cingulate cortex	8	4	4.9	1.7[c]
Orbitofrontal cortex	7	4	2.9	1.2[d]
Prefrontal cortex	7	1	1.7	0.9
Hippocampus	138	44	28.5[c]	11.9[c]
Putamen	20	3	3.4	1.7
Caudate nucleus	48	5	2.0[d]	2.0[d]
Amygdala	82	32	4.7[c]	4.3[c]
Lateral ventricles	35	15	14	3.5[c]
Third ventricle	27	9	6.8	2.5[c]
Gray matter	21	5	2.0[d]	1.2[c]
White matter	14	1	0.8	0.7
Globus pallidus	6	2	0.5[d]	0.4[d]
Thalamus	10	3	0.5[c]	0.5[c]
Temporal lobe	24	6	1.6[c]	1.2[c]
Vermal lobules	22	8	2.6[c]	1.4[c]

[a] Based on the assumption that the plausible effect is the one seen in each meta-analysis of each particular brain structure and condition.
[b] Based on the assumption that the plausible effect is half of what is seen in each meta-analysis.
[c] Statistically significant difference between expected and observed even after Bonferroni adjustment for the total number of tested conditions or brain structures.
[d] Nominally statistically significant difference between expected and observed.

tive to positive to achieve the same bias as if 10 times more negative studies were entirely suppressed. Selective analysis reporting bias emerges when there are many analyses that can be performed and only one of them, the one with the "best" results, is presented.[23,32] It is also facilitated when there are many steps in the analysis process that are subject to choices and measurements that can be biased. Some biases may also be facilitated when the assessors are not blinded[33] to whether each brain scan is coming from a case with the condition of interest or a control participant. Information about such quality safeguards is often not reported in the literature on brain volume abnormalities.

Although brain volume measurements are sophisticated, there is room for error. Magnetic resonance imaging measurements have average errors in the range of ±1.5%,[34,35] whereas changes of 5% may be introduced by scanner hardware or software.[34] Recognized sources of possible error include voxel misclassification during brain segmentation,[36,37] the partial volume problem (when volume is less than a voxel),[36,38,39] inconspicuousness of tissue edges, and head tilt. Nonsystematic errors will tend to dilute the observed effect sizes, if they are nondifferential. However, when errors are differential (eg, measurements are performed by observers who may favor a certain direction of the results), this can lead to inflated effects and spurious statistically significant associations. Moreover, nonsystematic mistakes may also occur during the analytic process.[40] Most studies that we assessed used magnetic resonance imaging. However, at least 1 meta-analysis[3] also included data on CT scans, and brain volume differences tended to be larger with CT measurements; this may be a manifestation of higher error rates in CT studies.[3] Finally, some structures are often possible to measure using various anatomical definitions.[2] Bias would be introduced if one assesses many different definitions and reports only the ones with the most significant results.

Brain volume differences may be confounded by drug treatment, inpatient vs outpatient status, differences in age and sex between the compared groups, severity of disease, and even disease definition per se. It is very difficult to control efficiently for all of these parameters. Selective analysis reporting may be introduced if investigators perform different analyses to adjust for various sets of confounders or exclude participants based on confounder values and then selectively report analyses based on the statistical significance of the results. There are many other aspects in the design, conduct, methods, analysis, and population characteristics of imaging studies that may affect the accuracy and reliability of the results. Case-control studies frequently report biased results because the cases are not representative of those affected in a population and/or because the controls are not representative of those not affected in the same population. It would be useful for experts in the field to adopt more unbiased designs, such as 2-stage sampling techniques.

The current evaluation did not consider voxel-based morphometry studies for which meta-analyses have started to appear recently.[41-43] Meta-analyses of such studies aim to reveal differences in gray matter density at specific brain coordinates rather than differences in volumes of prespecified regions of interest. These are whole-brain methods, and thus, in theory, they may avoid the selective reporting of selected regions of interest. However, the technique of activation likelihood estimation that is used for meta-analysis of voxel-based morphometry[41] has the disadvantage that it can use only data from studies that have significant differences between cases and controls. This strengthens the potential for significance chasing bias. A more recent method, signed differential mapping,[43] allows for the consideration of null findings and mitigates the excessive influence of single-study data sets. However, even if signed differential mapping allows for the incorporation of null findings, this will happen only if null findings are published so as to be available for meta-analysis.

Some limitations should be acknowledged in my study. First, several of the meta-analyses had substantial between-study heterogeneity, and the difference between O and E was larger in those meta-analyses with larger I^2 estimates. Heterogeneity may be a manifestation of bias affecting differentially the constituent data sets, but it may also reflect genuine differences across studies. It is possible that some of these effects are genuinely heterogeneous. However, with genuine heterogeneity, one would not necessarily expect that single-study data sets would pass a "desired" threshold of significance ($P < .05$) and yield so many statistically significant results. Empirical evidence from other fields where many associations are evaluated (eg, candidate gene associations) suggests that heterogeneity is often a marker of bias.[22] A better understanding of these genuine sources of diversity would require that we accumulate more unbiased data in the literature.

Second, even though the overall analysis suggests the presence of considerable bias, one cannot assume that all meta-analyses are equally affected. Probably the most useful application of the excess significance test is to give an overall impression about the average level of bias afflicting the field of brain volume abnormalities. The test can also be used to interpret separately the results of single meta-analyses, but here the interpretation should be more cautious. As shown in my results, some meta-analyses show strong evidence that they are affected by excess significance bias; some others seem spared of this bias, and their results can be considered to be reliable in this regard; and still others are difficult to interpret, mostly because of limited evidence. A negative test for excess significance in a single meta-analysis, especially one with few studies, does not exclude the potential for bias.

Third, the evaluation relied on effect sizes that had already been estimated in published meta-analyses because it would have been very inconvenient to perform 41 meta-analyses again from scratch. The data were obtained from meta-analyses published in high-profile peer-reviewed journals, but it is possible that some mistakes in the data may have been made, even in the best meta-analyses. However, there is no reason to believe that these would favor the presence or absence of excess significance bias. No effort was made to update these 41 meta-analyses with studies that appeared after the publication of each of the included meta-analyses. However, there is no reason to believe that these most recent studies would be much different in terms of susceptibility to bias. Moreover, the evaluation focused on meta-analyses published recently (ie, in the last 4 years).

Fourth, the exact estimation of excess significance can be influenced by the choice of plausible effect size, the potential miscalculation of the *P* values in the original data sets, and/or the miscalculation of power. This is why I have examined the influence of different effect sizes on the difference between O and E. Miscalculation of *P* values would require access to the raw data of each data set. For example, some of the excess significance may be in part due to *P* values in single data sets being reported as nominally significant owing to inappropriate assumptions (eg, equal variances). Power estimates with different assumptions (such as deviation from normality) and different software may diverge, but they are unlikely to change the big picture that almost all of these studies are small and largely underpowered and that the E is substantially smaller than the O.

In conclusion, the literature on brain volume differences is probably subject to considerable bias. This does not mean that none of the observed associations in the literature are true. It should be acknowledged that some meta-analyses may be more affected by bias than others and that some may be totally unbiased. However, the average level of bias is probably large, and steps should be taken to remedy the situation. Such steps could include, besides the use of newer technologies, the adoption of large multicenter studies, the standardization of definitions, outcomes, and analyses, and the registration of prespecified protocols for these studies. Large multicenter studies should be feasible, and it would be natural for such studies to use commonly agreed outcomes, definitions, and analyses in prespecified protocols that would be widely visible to all participating investigators and beyond. For most of the examined brain structures, definitions should be consistent, and this applies also to their subfields, which may yield additional insights if properly assessed.[44] Significance testing should not be used as a criterion for publication,[45,46] and journal editors can emphasize the need to make the full data (correlation matrices) and protocols available, as in other research fields.[47-49] After more than 25 years of research in this field, further progress requires stronger guarantees of reliability for the ensuing results.

Submitted for Publication: December 12, 2010; final revision received January 22, 2011; accepted January 28, 2011.
Published Online: April 4, 2011. doi:10.1001/archgenpsychiatry.2011.28
Correspondence: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, Stanford University School of Medicine, MSOB X306, 251 Campus Dr, Stanford, CA 94305 (jioannid@stanford.edu).
Author Contributions: Dr Ioannidis 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.
Financial Disclosure: None reported.

REFERENCES

1. Koolschijn PC, van Haren NE, Lensvelt-Mulders GJ, Hulshoff Pol HE, Kahn RS. Brain volume abnormalities in major depressive disorder: a meta-analysis of magnetic resonance imaging studies. *Hum Brain Mapp*. 2009;30(11):3719-3735.
2. Hajek T, Kopecek M, Kozeny J, Gunde E, Alda M, Höschl C. Amygdala volumes in mood disorders: meta-analysis of magnetic resonance volumetry studies. *J Affect Disord*. 2009;115(3):395-410.
3. Kempton MJ, Geddes JR, Ettinger U, Williams SC, Grasby PM. Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Arch Gen Psychiatry*. 2008;65(9):1017-1032.
4. Rotge JY, Guehl D, Dilharreguy B, Tignol J, Bioulac B, Allard M, Burbaud P, Aouizerate B. Meta-analysis of brain volume changes in obsessive-compulsive disorder. *Biol Psychiatry*. 2009;65(1):75-83.
5. Karl A, Schaefer M, Malta LS, Dörfel D, Rohleder N, Werner A. A meta-analysis of structural brain abnormalities in PTSD. *Neurosci Biobehav Rev*. 2006;30(7):1004-1031.
6. Stanfield AC, McIntosh AM, Spencer MD, Philip R, Gaur S, Lawrie SM. Towards a neuroanatomy of autism: a systematic review and meta-analysis of structural magnetic resonance imaging studies. *Eur Psychiatry*. 2008;23(4):289-299.
7. Steen RG, Mull C, McClure R, Hamer RM, Lieberman JA. Brain volume in first-episode schizophrenia: systematic review and meta-analysis of magnetic resonance imaging studies. *Br J Psychiatry*. 2006;188:510-518.
8. Boos HB, Aleman A, Cahn W, Hulshoff Pol H, Kahn RS. Brain volumes in relatives of patients with schizophrenia: a meta-analysis. *Arch Gen Psychiatry*. 2007;64(3):297-304.
9. Lawrie SM, Abukmeil SS. Brain abnormality in schizophrenia. A systematic and quantitative review of volumetric magnetic resonance imaging studies. *Br J Psychiatry*. 1998;172:110-120.
10. Videbech P, Ravnkilde B. Hippocampal volume and depression: a meta-analysis of MRI studies. *Am J Psychiatry*. 2004;161(11):1957-1966.
11. Honea R, Crow TJ, Passingham D, Mackay CE. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. *Am J Psychiatry*. 2005;162(12):2233-2245.
12. Wright IC, Rabe-Hesketh S, Woodruff PW, David AS, Murray RM, Bullmore ET. Meta-analysis of regional brain volumes in schizophrenia. *Am J Psychiatry*. 2000;157(1):16-25.
13. Bogerts B, Meertz E, Schönfeldt-Bausch R. Basal ganglia and limbic system pathology in schizophrenia: a morphometric study of brain volume and shrinkage. *Arch Gen Psychiatry*. 1985;42(8):784-791.
14. Husain MM, McDonald WM, Doraiswamy PM, Figiel GS, Na C, Escalona PR, Boyko OB, Nemeroff CB, Krishnan KR. A magnetic resonance imaging study of putamen nuclei in major depression. *Psychiatry Res*. 1991;40(2):95-99.
15. Bremner JD. Does stress damage the brain? *Biol Psychiatry*. 1999;45(7):797-805.
16. Rothstein H, Sutton AJ, Borenstein M. *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*. Chichester, England: Wiley; 2005.
17. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, Decullier E, Easterbrook PJ, Von Elm E, Gamble C, Ghersi D, Ioannidis JP, Simes J, Williamson PR. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*. 2008;3(8):e3081.
18. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4(3):245-253.
19. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med*. 2006;25(20):3443-3457.
20. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ*. 2006;333(7568):597-600.
21. Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *CMAJ*. 2007;176(8):1091-1096.
22. Kavvoura FK, McQueen MB, Khoury MJ, Tanzi RE, Bertram L, Ioannidis JP. Evaluation of the potential excess of statistically significant findings in published genetic association studies: application to Alzheimer's disease. *Am J Epidemiol*. 2008;168(8):855-865.
23. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648.
24. Cooper H, Hedges LV. *The Handbook of Research Synthesis*. New York, NY: Russell Sage Foundation; 1994.
25. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10:101-129. doi:10.2307/3001666.
26. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558.
27. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560.
28. Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JPA. Critical interpretation of Cochran's Q test depends on power and prior assumptions about heterogeneity. *Res Synth Methods*. 2010;1:149-161. doi:10.1002/jrsm.13.
29. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127(9):820-826.
30. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007;335(7626):914-916.
31. PS: Power and Sample Size Calculations, version 3.0, January 2009, developed 1997-2009 by William D. Dupont and Walton D. Plummer. Dept of Biostatistics, Vanderbilt University Web site. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize. Accessed February 23, 2011.
32. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*. 2008;358(3):252-260.
33. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
34. Day SJ, Altman DG. Statistics notes: blinding in clinical trials and other studies. *BMJ*. 2000;321(7259):504.
35. Howard MA, Roberts N, García-Fiñana M, Cowell PE. Volume estimation of prefrontal cortical subfields using MRI and stereology. *Brain Res Brain Res Protoc*. 2003;10(3):125-138.
36. MacFall JR, Payne ME, Krishnan KRR. MR Scanner geometry changes: phantom measurements compared with intracranial contents calculations. In: Proceedings of the International Society for Magnetic Resonance in Medicine; May 15-21, 2004; Kyoto, Japan. Abstract 2182.
37. Wang D, Chalk JB, Rose SE, de Zubicaray G, Cowin G, Galloway GJ, Barnes D, Spooner D, Doddrell DM, Semple J. MR image-based measurement of rates of change in volumes of brain structures: part II, application to a study of Alzheimer's disease and normal aging. *Magn Reson Imaging*. 2002;20(1):41-48.
38. Steen RG, Ogg RJ, Reddick WE, Kingsley PB. Age-related changes in the pediatric brain: quantitative MR evidence of maturational changes during adolescence. *AJNR Am J Neuroradiol*. 1997;18(5):819-828.
39. González Ballester MA, Zisserman A, Brady M. Segmentation and measurement of brain structures in MRI including confidence bounds. *Med Image Anal*. 2000;4(3):189-200.
40. Tofts PS, Barker GJ, Filippi M, Gawne-Cain M, Lai M. An oblique cylinder contrast-adjusted (OCCA) phantom to measure the accuracy of MRI brain lesion volume estimation schemes in multiple sclerosis. *Magn Reson Imaging*. 1997;15(2):183-192.
41. Haller JW, Banerjee A, Christensen GE, Gado M, Joshi S, Miller MI, Sheline Y, Vannier MW, Csernansky JG. Three-dimensional hippocampal MR morphometry with high-dimensional transformation of a neuroanatomic atlas. *Radiology*. 1997;202(2):504-510.
42. Fornito A, Yücel M, Patti J, Wood SJ, Pantelis C. Mapping grey matter reductions in schizophrenia: an anatomical likelihood estimation analysis of voxel-based morphometry studies. *Schizophr Res*. 2009;108(1-3):104-113.
43. Bora E, Fornito A, Yücel M, Pantelis C. Voxelwise meta-analysis of gray matter abnormalities in bipolar disorder. *Biol Psychiatry*. 2010;67(11):1097-1105.
44. Radua J, Mataix-Cols D. Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder. *Br J Psychiatry*. 2009;195(5):393-402.
45. Wang Z, Neylan TC, Mueller SG, Lenoci M, Truran D, Marmar CR, Weiner MW, Schuff N. Magnetic resonance imaging of hippocampal subfields in posttraumatic stress disorder. *Arch Gen Psychiatry*. 2010;67(3):296-303.
46. Ioannidis JPA. Journals should publish all "null" results and should sparingly publish "positive" results. *Cancer Epidemiol Biomarkers Prev*. 2006;15:185. doi:10.1158/1055-9965.EPI-05-0921.
47. Young NS, Ioannidis JP, Al-Ubaydli O. Why current publication practices may distort science. *PLoS Med*. 2008;5(10):e201.
48. Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, Cavalieri D, Gaasterland T, Hingamp P, Holstege F, Ringwald M, Spellman P, Stoeckert CJ Jr, Stewart JE, Taylor R, Brazma A, Quackenbush J; Microarray Gene Expression Data. The underlying principles of scientific publication. *Bioinformatics*. 2002;18(11):1409.
49. Baggerly K. Disclose all data in publications. *Nature*. 2010;467(7314):401.