The role of BDNF in depression : will the neurothrophin hypothesis sparkle on, long after the glitter of the firework is gone?

Molendijk, M.L.

CHAPTER 9

A systematic review on the association between val^{66}met and hippocampal volume – a genuine effect or a winners curse?

Molendijk ML
Bus BAA
Spinhoven P
Kaimatzoglou A
Oude Voshaar RC
Penninx BWJH
van IJzendoorn MH
Elzinga BM

Published as: A systematic review and meta-analysis on the association between BDNF val^{66}met and hippocampal volume – a genuine effect or a winners curse?
American Journal of Medical Genetics 2012; **159:** 731-740

SIGNIFICANCE: Over the years, the finding that met carriers at the BDNF val^{66}met locus have a relatively low hippocampal volume is a pillar under the neurotrophinn hypothesis to which basically all paper on the genotype refer (total number of citations for the first paper on this issue [Pezawas *et al.,* 2005] is > 500). Here however we show that this association probably is non-existent and due to underpowered studies.

ABSTRACT

Inconsistenties have been reported with regard to an association between val^{66}met, a single nucleotide polymorphism on the BDNF gene, and hippocampal volume. We performed a systematic review and a meta-analysis to determine the magnitude and direction of this putative association and estimated the potential influence of demographical, clinical, and methodological characteristics of studies. Tests of publication bias and time-related trends were performed and statistical power of the included studies was calculated. The literature search for MRI studies on differences in total hippocampal volume as a function of BDNF val^{66}met returned twenty-five records that fulfilled our criteria (total N = 3,620). Meta-analysis showed that carriers of a met allele had lower hippocampal volumes relative to val/val homozygotes (d = 0.13, P = .02). Between-study heterogeneity in effect size estimates was substantial and this could not be explained by demographical, clinical, and methodological differences across studies. Funnel plot inspection and trim-and-fill estimations suggested evidence for publication bias and effect sizes decreased substantially over the years (Pearson's r = -0.54, P < .01). All included studies were underpowered. This meta-analysis suggests that carriers of a met allele have lower total hippocampal volumes relative to val/val homozygotes. However, the effect sizes on this association converged closer to null with virtually each attempt at replication and were based on underpowered studies. Together our findings may suggest that the reported association between BDNF val^{66}met and hippocampal volume is not a genuine biological effect of the met allele but likely represents a winners-curse.

INTRODUCTION

Brain-Derived Neurotrophic Factor (BDNF) has been implicated in the pathophysiology of major depressive and bipolar disorder, and schizophrenia (Krishnan and Nestler, 2010). BDNF is a neuronal growth factor that has an array of functions including the induction of neuronal sprouting and differentiation (Poo, 2001). The role of BDNF is particularly evident in the hippocampus where it regulates processes such as learning (Lu and Gottschalk, 2000; Tapia-Arancibia *et al.,* 2008). Besides, by acting on hippocampal networks, BDNF is believed to be a moderator of mood (Taliaz *et al.,* 2009).

An intriguing feature of the expression of BDNF is that it is, unlike other neurotrophins, not only secreted constitutively but also in response to neuronal activity (*i.e.,* activity dependent secretion; Egan *et al.,* 2003). Interestingly, *in vitro* experiments have shown that the activity dependent secretion of BDNF is dependent on a single nucleotide site on the BDNF gene; val^{66}met, a valine into methionine insertion at codon 66 (Egan *et al.,* 2003; Chen *et al.,* 2004). Congruent with this is the finding that transgenic mice that carried a met allele had less dendritic complexity in the hippocampus and a reduced hippocampal volume (Magarinos *et al.,* 2010). Furthermore, in *in vitro* experiments the met allele has been linked to diminished neuronal integrity (Egan *et al.,* 2003; Stern *et al.,* 2010). Finally, human data are suggestive for an association between the met allele and the incidence of mood disorders (Verhagen *et al.,* 2008), schizophrenia (Gratacos *et al.,* 2007), and bipolar disorder (Rakofsky *et al.,* 2011). Taken together, these data add considerably to the idea that BDNF expression contributes to psychopathological characteristics and that this might be mediated by variation at the val^{66}met locus (for a critical review see Groves 2007).

In line with this idea, two high impact papers (Pezawas *et al.,* 2004; Szeszko *et al.,* 2005), using Magnetic Resonance Imaging (MRI) techniques, showed lower hippocampal volumes in carriers of a met allele relative to val/val homozygotes. This message inspired many and the association between val^{66}met and hippocampal volume became an area of interest, not at least because the hippocampus is considered to be a clinically relevant phenotype (MacQueen and Frodl, 2011). However, subsequent data with regard to this association is heterogeneous (*e.g.,* Dutt *et al.,* 2009). Reasons for this might be that the effect of val^{66}met on hippocampal volume is small and that therefore some studies may have lacked the necessary statistical power to detect it (*i.e.,* false negatives) or that the two pioneering papers may have overestimated the true effect (*i.e.,* a winners curse). Alternatively it could be that heterogeneity in findings is caused by demographical, clinical, or methodological differences across studies.

When faced with non-uniform findings it is useful to aggregate data over studies in order to learn about the most plausible nature of an association (Lohmueller *et al.,* 2003). Hence, we determined the magnitude of the putative association between BDNF val^{66}met and total hippocampal volume by means of a meta-analysis. The potential moderating influence of demographical, clinical, and methodological characteristics of studies were also tested and tests of publication bias were performed.

METHOD

Search Strategy

Using the terms: val^{66}met OR rs6265 AND hippocampus and val^{66}met OR rs6265 AND hippocampal volume two of us (BB and MM) searched the database PUBMED (www.ncbi.nlm.nih) through February 1st 2012 for human MRI studies on differences in hippocampal volume as a function of BDNF val^{66}met. The digital search was supplemented by a backward search in which all the references that were made to the 2 seminal papers were screened and by examining the reference sections of the retrieved records. We selected for inclusion human MRI studies that reported on differences in total hippocampal volume between val/val homozygotes and carriers of a met allele. Inclusion was independent of demographic (*e.g.,* gender), clinical (*e.g.,* diagnostic status), and methodological characteristics (*e.g.,* Voxel-Based Morphology

[VBM] or actual volume measurements) of the sample and the study. Our search yielded 81 papers of which 25 records (*k*) fulfilled our inclusion criteria (total *N* = 3,620). For detailed information on the search strategy and the results of this strategy we refer to the flow chart (**Figure 1 ↓**).

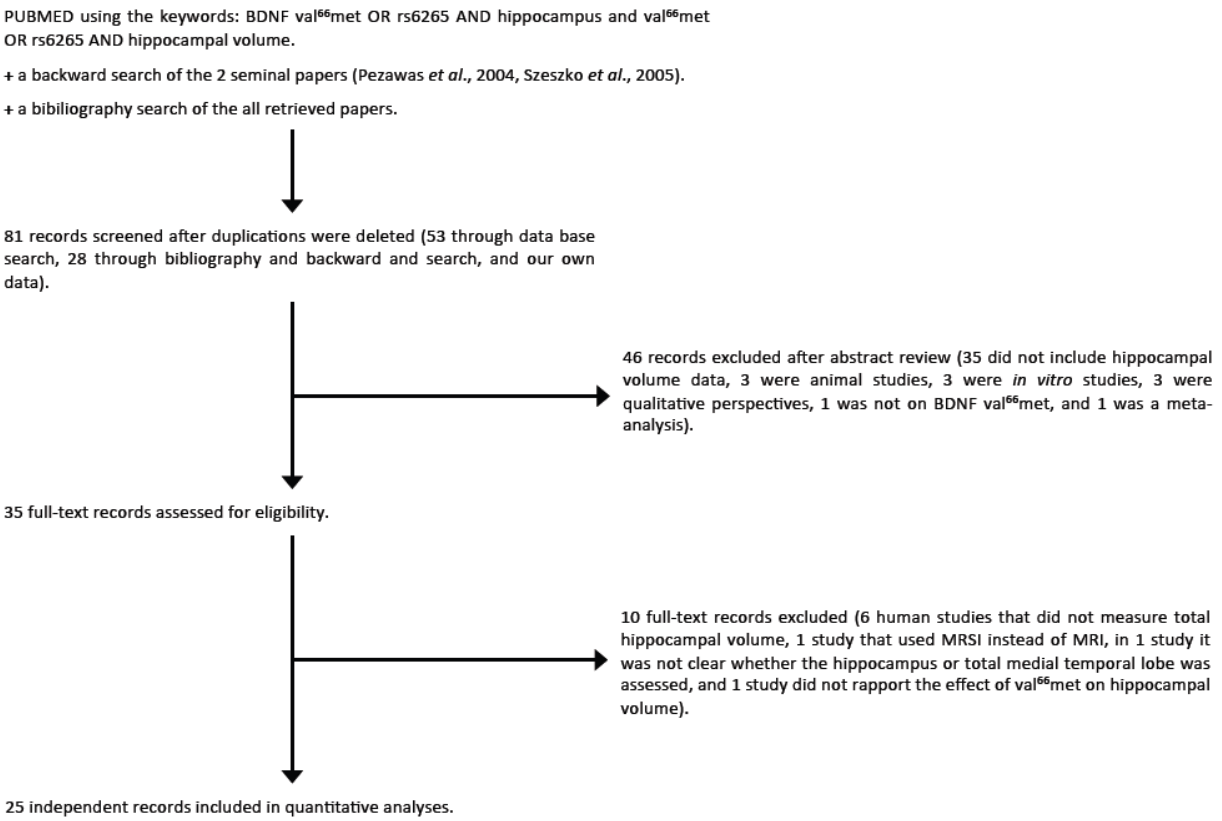

Figure 1. Flow-chart of the search for papers on the association of val^{66}met with hippocampal volume

Data Extraction

We extracted mean total hippocampal volume and Standard Deviation (*SD*) (or *t* and *P* values and the direction of the effect) as a function of BDNF val^{66}met genotype. These outcomes were weighted using inverse variance methods (Borenstein *et al.,* 2009) and converted to standardized Cohen's *d* metrics (Cohen, 1988). Here, a positive value of this metric indicated larger hippocampal volumes in val/val homozygotes relative to carriers of a met allele. In those cases where non-significant results were reported without the necessary statistics to calculate Cohen's *d* (2 records: Karnik *et al.,* 2010; Gerritsen *et al.,* 2011), we assigned the strength of the difference between val/val and carriers of a met allele in hippocampal volume an estimated effect size of 0. Where non-significant results were reported with sufficient information to calculate an effect size, but not the direction of the effect (1 record: Agartz *et al.,* 2006), we assigned the association a Cohen's *d* that was, with regard to its direction, concordant with the study hypothesis. To indicate whether effect size imputation was associated to different effect sizes, we constructed a binary variable indicating whether imputation had taken place. In a meta-regression framework this variable was tested for association with Cohen's *d*. Two of the included records reported longitudinal data (Koolschijn *et al.,* 2010; Millan Sanchez *et al.,* 2012). We included the baseline data of these studies since more subjects were available at baseline compared to follow-up.

In addition to hippocampal volumes and genotype, we extracted data on (**I**) demographical characteristics: mean age, percentage females, ethnicity, Minor Allele Frequency (MAF), and genotype

frequencies of the sample; (**II**) clinical characteristics: psychiatric status (i.e., percentage of the sample with current depression, schizophrenia, or bipolar disorder and the percentage healthy controls of the sample) and psychotrophic medication use (percentage of the sample that used antidepressants, antipsychotics and/or mood stabilizers); and (**III**) methodological characteristics of the study: method of hippocampal volume extraction (VBM versus actual volume measurements), Hardy-Weinberg equilibrium, and whether the hippocampus was traced manually or automatically.

Quality Assessment

We used the criteria set forth by the Strengthening Reporting of Genetic Association Studies (STREGA; Little *et al.,* 2009) and the Strengthening Reporting of Observational Studies in Epidemiology (STROBE; von Elme *et al.,* 2007) checklists using the 11-item list adaption from Karg *et al.* (2011) to evaluate the methodological quality of the included studies. Overall quality score was defined as the frequency of relevant criteria that were met by each individual study. Independent quality assessments were performed by AK and MM. Agreement among the raters proved to be excellent (Cohen's Kappa=0.83, Standard Error (*SE*) = 0.04). Overall, the quality of the included studies was good (mean = 0.86, *SD* = 0.14, range 0.56 - 1.00). Quality ratings of the studies are presented in **Table S1** in **Appendix IV** of this thesis.

Statistical analysis

Meta-analytical calculations were performed using Comprehensive Meta-Analyses version 2.0 (CMA 2.0; Borenstein *et al.,* 2009) with statistical significance set at $P < .05$.

A random effects model was applied to calculate Cohen's *d* (± 95% Confidence Interval (*CI*)) on the difference in total hippocampal volume between val/val homozygotes and carriers of a met allele. Heterogeneity between studies was assessed using the *Q* statistic (Borenstein *et al.,* 2009). Given the possible impact of psychiatric diagnoses on hippocampal volume (MacQueen and Frodl, 2011) meta-analyses and heterogeneity assessments were repeated stratified by psychiatric diagnosis (no diagnosis versus any diagnosis). The difference in effect-sizes that were acquired in these analyses was assessed using a *z* difference statistic.

In a series of meta-regression analyses the possible moderating effects of demographical, clinical and methodological differences across studies on Cohen's *d* were evaluated. The first of these analyses was carried out to test the effects of demographical and methodological characteristics and was run using the data from all included studies. In addition, we tested the clinical characteristic: healthy controls versus any disorder in this analysis. In a second analysis the moderating effects of the demographical and methodological characteristics were assessed using the data from healthy control samples only. This was done to exclude the noise that might have been caused by diagnostic or psychotrophic treatment status of the patient samples. A third analysis was conducted in patient samples to specifically test the moderating effects of psychiatric status (depression versus no depression, schizophrenia versus no schizophrenia, and bipolar disorder versus no bipolar disorder) and psychotrophic medication use (yes versus no). In case of >1 statistically significant moderator, meta-regression analyses were followed up by multivariable regression analysis (SPSS Inc, Chicago, Ill) including the significant moderators in order to learn about their relative contributions to Cohen's *d*.

Publication bias was assessed by funnel plot asymmetry inspection and the Egger test (Egger *et al.,* 1997). In case of publication bias, a trim-and-fill procedure was performed. The trim-and-fill procedure is a procedure that provides an estimation of the effect size after potential bias has been taken into account (Duval and Tweedie, 2000; Peters *et al.,* 2007). Tests of time-related trends were performed by correlating year of publication with weighted Cohen's *d*. Time-related trends were visualized by means of a cumulative

meta-analysis (a meta-analysis that calculates an aggregated effect size for each study that is added to the literature) and scatter-plots. *A posteriori* power and sample size calculations were performed using G*Power (Faul *et al.,* 2009).

RESULTS

Description of samples

The number of subjects of the included studies ranged from *n* = 34 to *n* = 572 (mean = 145, *SD* = 122). In 14 out of the 25 studies (56%) the majority of subjects was female. Mean age of the samples ranged from 23 years to 72 years (mean = 40, *SD* = 14). Eleven of the 25 included studies (44%) reported data on healthy subjects only (*n* = 1,784). The remaining 14 studies (56%) reported data on both healthy subjects (14 subsamples, *n* = 981) and patients with a diagnosis of a psychiatric ilness (depression [7 subsamples, *n* = 431], Schizophrenia or psychosis [6 subsamples, *n* = 345], bipolar disorder [2 subsamples, *n* = 50], and anxiety [1 subsample, *n* = 29]). Some studies (Benjamin *et al.,* 2006 and Gruber *et al.,* 2011) did not provide sufficient information to calculate Cohen's *d* separately for the healthy and the patient samples. Because of this, these studies were not included in the stratified meta-analyses and meta-regression analyses. The numbers in these analyses, therefore, do not add up to the total of *N* = 3,620. **Table 1 ↓** shows basic information on the included records.

Meta-analysis

The results of the meta-analysis over all studies (*k* = 25, *N* = 3,620) showed that carriers of a met allele had lower hippocampal volumes as compared to val/val homozygotes (*d* = 0.13 ± 0.06, 95% *CI* = 0.03 to 0.24, *z* = 2.41, *P* = .02; see **Figure 2 ↓**, panel A for a forest-plot). Analyses stratified by psychiatric diagnosis (no diagnosis versus any diagnosis) revealed similar point estimates for non-patient (*d* = 0.16 ± 0.06, 95% *CI* = 0.04 to 0.27, *z* = 2.57, *P* = .01, *k* = 23, *n* = 2,542) and patient samples (*d* = 0.17 ± 0.11, 95% CI= -0.05 to 0.38, *z* = 1.54, *P* = .12, *k* = 12, *n* = 692)($P_{difference}$ = .96). Substantial heterogeneity across studies was identified in the analyses that were run on the data of all samples (*Q* = 54.47, *P* < .001) on the data of healthy samples (*Q* = 35.11, *P* < .05), and on the data of psychiatric samples (*Q* = 18.11, *P* = .08).

Table 1. Summary of characteristics of studies measuring total hippocampal volume differences between val/val homozygotes and carriers of a met allele at the val[66]met locus presented by year of publication

Author, year	*N*	% female	age	ethnicity	% of *N*	MAF	patient status	% of *N*
Pezawas *et al*., 2004	111	50%	34 [a]	Caucasian	100%	*NK*	healthy controls	100%
Szeszko *et al*., 2005	44	45%	27	Caucasian	100%	0.19	healthy controls schizophrenia	67% 43%
Agartz *et al*., 2006	101	30%	42 [a]	Caucasian	100%	0.19 [a]	healthy controls schizophrenia	51% 49%
Bueller *et al*., 2006	36	61%	27	Caucasian African American Asian	67% 19% 14%	0.21	healthy controls	100%
Frodl *et al*., 2007	120	48%	43	*NK*	*NK*	0.19	healthy controls depression	50% 50%
Miyajima *et al*., 2008	61	68% [a]	63 [a]	Caucasian	100%	0.19 [a]	healthy controls	100%
Takahashi *et al*., 2008	62	40%	25	Asian	100%	0.39	healthy controls schizophrenia	53% 47%
Chepenik *et al*., 2009	34	53%	*NK*	Caucasian African American Other [b]	82% 8% 10%	0.20	healthy controls bipolar disorder	47% 53%
Dutt *et al*., 2009	383	50%	43	Caucasian	100%	*NK*	healthy controls unaffected relatives psychosis	16% 50% 33%
Gatt *et al*., 2009	89	51% [a]	36 [a]	Caucasian	100%	0.20 [a]	healthy controls	
Jessen *et al*., 2009	163	56%	43	Caucasian	100%	*NK*	healthy controls depression	48% 52%
Joffe *et al*., 2009	113	48% [a]	37 [a]	Caucasian	100%	0.21	healthy controls	100%
Schofield *et al*., 2009	161	47%	32	*NK*	*NK*	0.22	healthy controls	100%
Toro *et al*., 2009	331	52%	*NK*	Caucasian	100%	0.20	healthy controls	100%
Benjamin *et al*., 2010	173	65% [a]	69 [a]	Caucasian	100%	*NK*	healthy controls depression	67% 33%
Karnik *et al*., 2010	129	54%	49	Caucasian African American	90% 10%	0.17	healthy controls	100%
Koolschijn *et al*., 2010	177	28%	37	Caucasian	100%	0.20	healthy controls schizophrenia	51% 49%
Cole *et al*., 2011	188	55%	40	Caucasian	100%	0.22	healthy controls depression	59% 41%
Gerritsen *et al*., 2011	572	63%	23	Caucasian	100%	0.23	healthy controls	100%
Gonul *et al*., 2011	73	66%	32	Caucasian	100%	0.25	healthy controls depression	55% 45%
Gruber *et al*., 2011	105	47%	38	Caucasian	100%	0.23	healthy controls schizophrenia bipolar disorder	37% 32% 30%
Kanellopoulos *et al*., 2011	56	63%	72	Caucasian	100%	0.27	healthy controls depression	59% 41%
Richter *et al*., 2011	138	67%	25	Caucasian	100%	0.27	healthy controls	100%
Milan Sanchez *et al*., 2012	43	7% [a]	57 [a]	Caucasian	100%	0.24 [a]	healthy controls	100%
Molendijk *et al*., 2012	157	67%	37	Caucasian	100%	0.18	healthy controls depression [c, d] anxiety [d]	20% 61% 19%

Abbreviations: MAF, Minor Allele Frequency; *NK*, Not Known; e-pub, e-pub ahead of print.
[a] Estimated from larger sample
[b] Not further specified ethnicity, but not Caucasian, African American, or Asian
[c] Included a diagnosis of depressive disorder (*n* = 43, 45%) or comorbid depressive/anxiety disorder (*n* = 52, 55%)
[d] Included a diagnosis of social phobia, panic disorder, generalized anxiety disorder, and/or agoraphobia

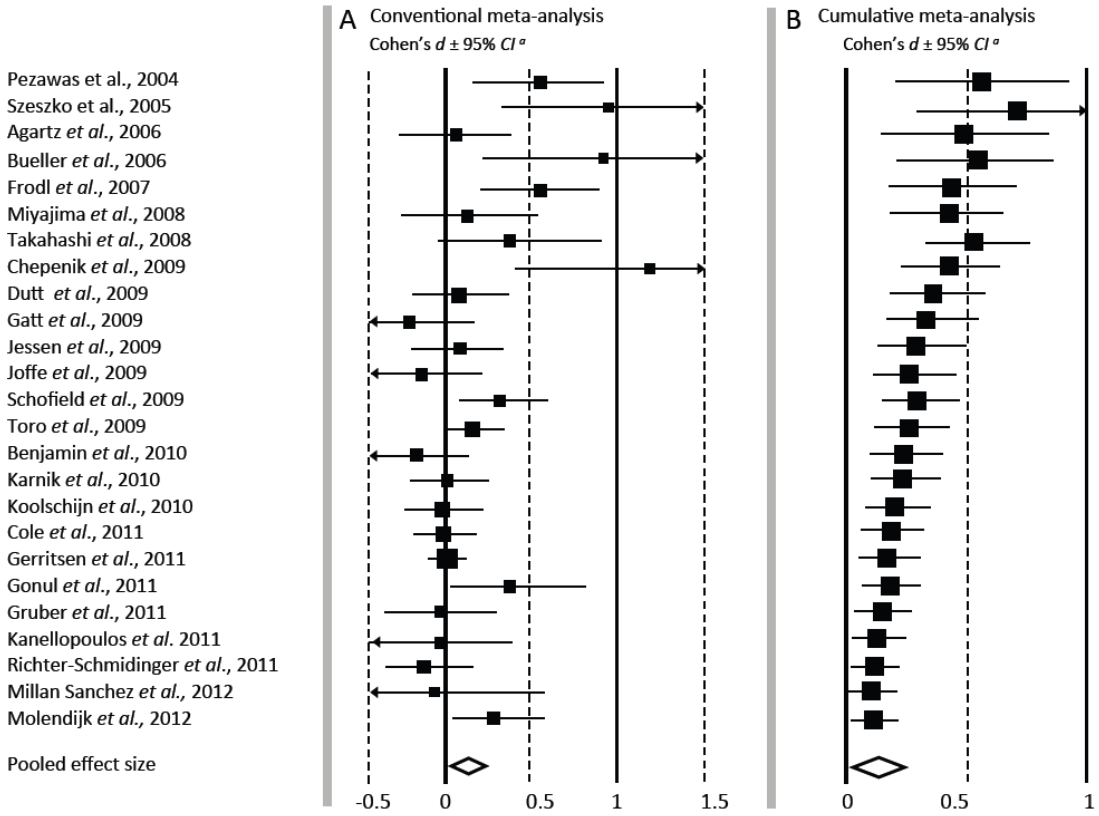

Figure 2. Forrest plot of a conventional meta-analysis (panel **A**, left side of the Figure) and a cumulative meta-analysis (panel **B**, right side of the Figure)

Meta-regression analysis

We evaluated the potential moderating effects of demographical, clinical, and methodological differences across studies in a series of 3 meta-regression analyses. Analyses were conducted separately using the data from all included studies, using the data from healthy control samples only, and using the data from patient samples only. **Table 2** ↓ provides the coefficients that were obtained in these analyses. In sum, mean age of the sample explained a significant amount of variance in weighted *d* (r = -0.43, R^2 = 0.18, P < .05), but most pronounced in the data that were derived from healthy samples. This effect was such that effect sizes were lower in healthy samples in which the subjects were older. Effects of other demographical, clinical, and methodological moderators were not observed. Imputation of effect size (3 records: Agartz *et al.,* 2006; Karnik *et al.,* 2010; Gerritsen *et al.,* 2011) and methodological quality of the included studies also were unrelated to weighted effect size.

Publication bias, and time-related trends, and Sample size calculations

Visual inspection of the funnel plot suggested evidence for publication bias. Egger's test confirmed this (Egger's Intercept = 1.71, 95% *CI* = 0.16 to 3.26, t = 2.29, P = .02). A trim-and-fill estimation suggested that the addition of 2 small and non-significant studies that had to be trimmed and filled would be sufficient to result in a non-significant agregated Cohen's *d* (random effects model) of 0.09 (95% *CI* = -0.02 to 0.22, not statistically significant; see **Figure 3** ↓ for the funnel-plot with observed and imputed values).

Test of time-related trends showed a significant correlation between year of publication (2004 to 2012) and Cohen's *d* (r = -0.54, R^2 = 0.29, P < .01). This effect was consistently found in healthy control samples (r = -0.49, R^2 = 0.24), patient samples (r = -0.55, R^2 = 0.30), and mixed healthy–patient samples (r = -0.55, R^2 = 0.30). The observation that effect sizes decreased over the years is illustrated in **Figure 1**, panel B ↑ (a

cumulative meta-analysis) and in **Figure 4 ↓** (a scatter-plot on the relation between year of publication and effect size).

Table 2. Correlations of demographical, clinical, and methodological study characteristics with Cohen's *d* on the relation between val[66]met and hippocampal volume

	All samples (*k* = 25, *N* = 3,620)	HC samples (*k* = 23, *n* = 2,542) [a]	Patient samples (*k* = 12, *n* = 692) [a]
Demographical/Study characteristics			
Gender (percentage female)	0.04	-0.01	-0.19
Age (mean, years)	-0.36	-0.43* [b]	-0.35
Ethnicity (1 = mixed, 2 = Caucasian)	-0.29	-0.33	-0.40
Minor allele frequency	-0.08	-0.08	-0.14
Sample size	-0.31	-0.31	-0.40
Study quality (frequency of criteria met)	0.12	-0.04	0.45
Clinical characteristic			
Psychiatric diagnosis (1 = no, 2 = yes)	0.18	*NA*	*NA*
Major depressive disorder (1 = no, 2 = yes)	*NA*	*NA*	-0.29
Bipolar disorder (1 = no, 2 = yes)	*NA*	*NA*	0.39
Schizophrenia (1 = no, 2 = yes)	*NA*	*NA*	0.15
Psychotropic drugs (1 = no, 2 = yes)	*NA*	*NA*	-0.22
Methodological characteristics			
VBM (1 = no, 2 = yes)	0.12	0.24	0.01
Magnetic strength (1 = 1.5 Tesla, 2 = 3 Tesla) [c]	-0.09	-0.13	-0.07
Manual hippocampal measurement (1 = no, 2 = yes)	0.31	0.28	0.20

Abbreviations: HC, Healthy Control; *NA*, Not Applicable; VBM, Voxel Based Morphology.
[a] Note. Numbers do not add up to the total *N* of 3,620. This is because some studies (Benjamin *et al.,* 2006 and Gruber *et al.,* 2011) did not provide sufficient information to calculate Cohen's *d* separately for the healthy sample and the patient sample.
[b] Mean age did not remain a statistically significant predictor of Cohen's *d* in a multivariable regression analysis in which year of publication also was added as a predictor variable, whereas the latter did.
[c] One study measured at 1 Tesla [Toro *et al.,* 2009] and was coded as 1.5 Tesla. Excluding this study from analysis did not change the results.
* denotes statistical significance at *P* < .05

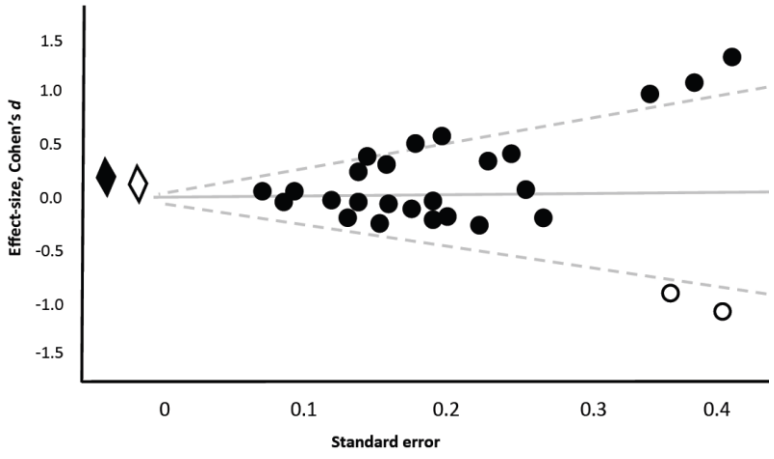

Figure 3. Funnel plot and trim-and-fill estimation showing the typical pattern of publication bias. Filled and open data points depict observed and imputed values respectively. The filed diamond depicts the aggregated point estimate (*d* = 0.13, *P* = .02) and the open diamond the aggregated point estimate after imputation of two studies (*d* = 0.09, not statistically significant).

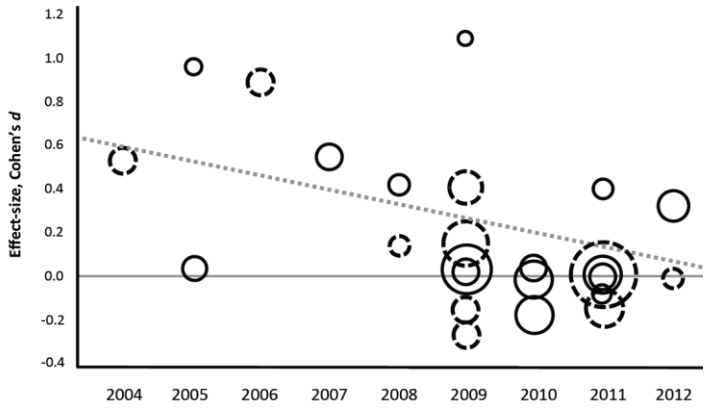

Figure 4. Scatter plot showing the relation between year of publication and standardized Cohen's *d* (weighted by the inverse of the variance) on the association of val^{66}met and total hippocampal volume (Pearson's *r* = -0.54, *P* < .01). Dashed bordered circles indicate studies that included healthy subjects only (*r* = -0.49). Solid bordered circles indicate studies that included both healthy control subjects and patients (*i.e.,* depression, schizophrenia, and bipolar disorder) (*r* = -0.55).

Given that year of publication and age both were significantly associated with effect size these variables were analyzed together in a multivariable regression model. Results of this analysis showed that the effect of year of publication on weighted *d* remained statistically significant (*B* = -0.07, 95% *CI* = -0.14 to -0.01, β = -0.43, *P* = .03) whereas the effect of age disappeared (*B* = -0.01, 95% *CI* = -.02 to 0.01, β = -.35, *P* = .08). It should be noted though that the multivariable statistics should be interpreted with caution when using meta-analysis data because the risk of over-fitting and spurious results (Sterne *et al.,* 2001). Also in our data, if we corrected the standard error for the use of meta-analytic data this relation lost its statistical significance. Notwithstanding this, through data inspection we recognized that the negative association between mean age and Cohen's *d* that we observed in univariable tests, might have been driven by null associations in 2 recently published studies in samples with the relatively high mean ages (~ 70 years; Benjamin *et al.,* 2010; Kanellopoulos *et al.,* 2011). Indeed, if these studies were excluded from the meta-regression, the effect of year of publication remained similar (*r* = -0.47, *P* < .05) whereas the effect of mean age of the sample lost its significance (*r* = -0.29, *P* = .17)

Based on the aggregated effect size we calculated the sample size that is needed to detect a relation between variation at val^{66}met (with the MAF being 0.25) and total hippocampal volume with a power of 0.80 at an α-level of .05. This calculation suggested that 1,900 subjects (1,086 val/val homozygotes and 814 carriers of a met alle) would be neccesary to detect an association of the met allele with total hippocampal volume. Statistical power of the included studies ranged from ~ .07 for the study with the smallest sample size (Chepenik *et al.,* 2009 [*n* = 34] reported effect size *d* = 1.20) to ~ 0.30 for the largest sample size (Gerritsen *et al.,* 2011 [*n* = 572] reported effect size *d* ~ 0). Thus, all the included studies were underpowered.

DISCUSSION

The main goal of this paper was to determine, by meta-analysis, the magnitude and direction of the relation between BDNF val^{66}met and hippocampal volume. Our results, based on 25 samples and a total of 3,620 subjects, suggest that carriers of a met allele have slightly lower total hippocampal volumes (*d* = 0.13) relative to val/val homozygotes. This finding has a plausible biological basis as it can be derived by the findings that BDNF regulates the sprouting and survival of neurons in the hippocampus (Lu and Gottschalk, 2000) and that the met allele is associated with abnormal activity of BDNF in hippocampal neurons (Egan *et al.,* 2003). Hence, the lower hippocampal volume in met carriers is mediated through aberrant trophic support by BDNF. Notwithstanding meta-analytical significance and concordance with biological knowledge, several outcomes of the meta-analyses indicate that the lower hippocampal volume in met carriers is not a genuine biological effect of the met allele but likely has an artificial basis.

Between-study heterogeneity in outcomes in genetic imaging studies may, in general, be due to associations that exist in some populations but not in others or might stem from between-study differences in methodology. Given heterogeneity in a number of characteristics across the studies, it may not be surprising that the reported effect sizes were variable as well (*i.e.,* 7 positive and statistically significant studies and 18 statistically inconclusive positive and negative studies). Through stratified meta-analyses and meta-regression analyses we aimed to identify the sources of this heterogeneity. This is an important venue to pursue as identifying factors that explain variance in outcomes may hint to possible mechanisms that thrive an association. Both types of analyses, however, gave little reason to suspect that heterogeneity in demographical, clinical, and methodological characteristics across studies was systematically related to heterogeneity in effect-sizes. Specifically, we would like to add that manual versus automatic hippocampal volume measurements and the use of 1.5 Tesla versus 3 Tesla also were not associated, structurally, with differences in effect-sizes. An evaluation of the relation between methodological quality of each of the included studies and imputation of effect size similarly showed no relation with Cohen's *d*. This lack of association is an important observation because it justifies the broad set of inclusion criteria that was applied here. However, it should be noted that the use of meta-regression analysis might be hazardous with regard to the occurrence of false positive and negative findings because the number of data-points on which the results of these analyse are based ussualy is rather small (*i.e.,* the number of studies that are included in a meta-analysis; Munafo and Flint, 2004).

Between-study heterogeneity may have artificial sources as well. We detected two such sources in the aggregated data set. First, we consistently observed, over the mixed healthy-patient samples, healthy samples, and patient samples that effect sizes converged closer to null with virtually each subsequent attempt at replication. Second, clear evidence for publication bias was identified. Publication bias typically results from negative studies that are left unpublished and/or from selective outcome reporting (Ioannidis 2011). Together, the decrease in effect size estimates over the years and the publication bias suggest that the observed aggregated effect size (d = 0.13) is an overestimation for the true or most plausible effect size on the association of interest.

Yet another finding from this meta-analysis is that the studies included in our meta-analysis were all underpowered. In fact, *a posteriori* power calculations revealed that the power of the included studies ranged from as low as ~ 0.07 to only ~ 0.30 to detect an effect of the met allele on hippocampal volume. Given that a low level of power increases the ratio of false to true positives (Sterne and Davey Smith, 2001), it seems likely that some false positive findings were among the studies that we included. It should be noted that evidence for increases in the ratio of false to true positives could not directly be extracted from our aggregated data. We did, however, find moderate negative correlation coefficients (albeit non-significant) for the relation between the number of subject in a study and effect size (range: -0.31 to -0.40), that is, larger samples tended to yield smaller effects.

There are some limitations with regard to the methods that were used to detect publication bias. A core problem with regard to the interpretation of funnel-plot asymmetry is that one never knows whether the funnel plot asymmetry is truly due to publication bias or whether it is due to unmeasured differences between studies (Munafo and Flint, 2004). Indeed, heterogeneity in effect-sizes may have come from sources that were not tested in our study, such as the duration or dose of psychotrophic medication use, disease severity, and exposure to stress (MacQueen and Frodl, 2010). Also the key assumption of the trim-and-fill method that the most extreme effect sizes are the ones that are left unpublished has been questioned. However, simulation experiments have shown superiority of the trim-and-fill method above other methods to quantify publication bias when between-study heterogeneity in outcomes is present (Peters *et al.,* 2007).

A limitation that we would like to add is that we could not test the hypothesis of differences in hippocampal volumes between subjects who were homozygous for the met allele (*i.e.,* met/met) and heterozygote val/met subjects, that is a potential dose-dependent effect of the met allele. The frequency of occurrence of the met/met variant is particularly low, at least in Caucasian samples (Petryshen *et al.,* 2011), and none of the included studies reported outcome estimates for this particular variant. Related, the majority of subjects in the studies that were included in our analyses were Caucasian from origin, except for 1 study that reported a positive non-significant effect in an Asian sample (Takahaski *et al.,* 2008). Thus our results might be less applicable for subjects who are of ethnic backgrounds other than Caucasian. In addition, it could be that variation at the val^{66}met locus of the BDNF gene is important for hippocampal morphology only in interaction with childhood trauma exposure for which some evidence exists (Gatt *et al.,* 2009) although 2 of the in this meta-analysis included records (Gerritsen *et al.,* 2011; Molendijk *et al.,* 2012) could not replicate this phenomenon. Another limitation of our study might be that we focused on total hippocampal volume whereas the morphology of the hippocampus is complex and can, for example, be subdivided in a head, a body and a tail (Maller *et al.,* 2007). It could be that the effect of the met allele is limited to morphologically specific sites of the hippocampus (see Montag *et al.,* 2009).

In sum, we carried out a systematic review and meta-analysis on the association between val^{66}met and total hippocampal volume. The results that are reported here indicate that carrying a met allele at the BDNF val^{66}met locus is associated with lower hippocampal volumes. So, one might conclude that the met allele has an effect on hippocampal morphology. However, we observed that effect size estimates converged closer to null with virtually each attempt at replication and that all studies on the subject matter were largely underpowered. Furthermore, we found evidence for publication bias inflating the association reported in the literature. Altogether, this not only suggests that the effect is inherently smaller than reported here (d = 0.13) but also calls into question whether the observed effect is a biological effect of the met allele or whether it is an artifact of underpowered studies. We therefore conclude that variation at the BDNF val^{66}met locus is not likely to account for individual differences in hippocampal volume but rather that the association is subject to a *winners curse*, with large effect sizes found in a few early studies and increasingly smaller effect sizes in later (better-powered) studies.