Published in final edited form as:

Cortex. 2016 January ; 74: 338–342. doi:10.1016/j.cortex.2015.02.019.

Failed replications, contributing factors and careful interpretations: commentary on “A purely confirmatory replication study of structural brain-behaviour correlations” by Boekel et al., 2015

Nils Muhlert¹ and Gerard R. Ridgway^{2,3}

¹School of Psychology and Cardiff University Brain Research Imaging Centre, Cardiff University, UK

²FMRIB Centre, Nuffield Department of Clinical Neurosciences, University of Oxford, UK

³Wellcome Trust Centre for Neuroimaging, UCL Institute of Neurology, London, UK

Keywords

Replications; VBM; Brain structure

The structure of the human brain reflects a wealth of functionally relevant information, and is influenced by both genes and environment. Gray matter volumes both at global and regional levels are highly heritable (den Braber et al., 2013; Posthuma et al., 2002; Winkler et al., 2010; Wright, Sham, Murray, Weinberger, & Bullmore, 2002), and map clearly onto specific genotypes (Durstun et al., 2005; Toga & Thompson, 2005). Regional brain volumes are typically stable over time (den Braber et al., 2013) but can increase following new learning (e.g. Woollett & Maguire, 2011), demonstrating experience-dependent plasticity well into adulthood. In addition, brain structure alterations resulting from neurodegenerative disease are linked to concomitant changes in functions associated with those structures, such as memory problems associated with medial temporal lobe atrophy (e.g. Leube et al., 2008). These findings provide convincing evidence that differences in brain structure can reflect behaviour.

Recent studies have moved from examining how group differences in brain structure reflect those in behaviour, to assessing how subtle inter-individual differences in brain structure map on to normal variation in behaviour (Kanai & Rees, 2011). Such studies have linked anatomical structure to personality traits (e.g. DeYoung et al., 2010) and to impulsivity traits (e.g. Matsuo et al., 2009). Extreme scores on these traits are often observed in those with psychopathology, such as unipolar and bipolar depression (Johnson, Carver, & Joormann, 2013; Zapolski, Guller, & Smith, 2012), and so these findings support a continuum view of the risk factors and potential biomarkers for psychiatric conditions.

Whilst the number of studies linking brain-structure to behaviour has rapidly increased over the last decade, a parallel set of studies have highlighted how methodological choices can influence these findings. These critiques can be separated loosely into those demonstrating the confounding influence of acquisition protocols, such as scan parameters (Tardif, Collins, & Pike, 2009) and field strength (Tardif, Collins, & Pike, 2010), the influence of pre-processing choices, such as smoothing kernel size (Shen & Sterr, 2013), masking (Ridgway et al., 2009) or the use of modulation (Radua, Canales-Rodriguez, Pomarol-Clotet, & Salvador, 2014), and the influence of particular statistical analyses (Rajagopalan, Yue, & Pioro, 2014). This has led to guidance on how to report anatomical imaging studies, to ensure they are easily interpretable and repeatable (Ridgway et al., 2008). Importantly, these methodological papers do not question the validity of the approach (although see Bookstein, 2001). Indeed, studies comparing voxel-based morphometry (VBM), the most widely used method for automated brain volume measurement, to manually defined region-of-interest approaches find highly similar results, indicating convergent validity (Bergouignan et al., 2009).

A recent study, however, calls into question the strength of some structural brain-behaviour correlations. In an attempt to replicate 17 separate findings from five published MRI studies, Boekel et al. (2015) found that, aside from one replicated effect, their data provided only weak evidence for them and in most cases favoured the null hypotheses, with “moderate to strong evidence” that no associations exist for 8 of the 17 effects. Effect size estimates were also found to be substantially lower than those from the original studies. Here we consider the methods and implications of the Boekel study, beginning with four general comments on the study as a whole, before focusing on the eight effects from two studies using VBM.

First, it has been suggested that convincing direct replications often require a much larger sample than the original study (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014), yet Boekel et al have smaller samples for 16 of the 17 effects (with n varying from 31 to 36). The consequent reduction in power was partially offset by the use of region-of-interest approaches (reducing the number of comparisons), one-tailed hypothesis testing (increasing the sensitivity) and Bayesian statistics. The Bayes factor may be less sensitive to sample size than classical frequentist statistical approaches as the probability of obtaining the alternative hypothesis remains fairly constant as sample size increases (Dienes, 2008). Nevertheless, an argument could be made that the 9/ 17 effects without moderate or strong evidence (in either direction) are not “failed replications”, but simply underpowered replication attempts. In particular, the first three effects from Westlye, Grydeland, Walhovd, and Fjell (2011), which had $n=132$ subjects, do not look convincingly refuted in Boekel’s Figure 8.

Second, although Boekel et al. cite papers on non-independence and double-dipping (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Vul, Harris, Winkielman, & Pashler, 2009), it is worth further highlighting that only two of the five studies report unbiased correlation estimates (via non-circular quantification of the effect size using a second data-set, with regions/peaks defined in the first data-set), shown in purple in Figure 8. The studies that did not carry out replications aimed primarily to detect and localise regional brain structure-behaviour associations, rather than to quantify the strength of any such association

(which would have entailed circular analyses). Since the methods used to detect these associations may bias the size of the effect readers should not “read too much into these estimates and overinterpret their value” (Kriegeskorte et al., 2010). Replications aimed at quantifying the size of such effects, such as is carried out for these three studies in the Boekel paper, will typically find reduced effects (Kriegeskorte et al., 2010), a point worth bearing in mind when interpreting these particular findings.

Third, an additional potential contributor to the observed shrinkage was the correction for nuisance variables used in the models. For all of the replications, structural measures had been “corrected for age and gender [sometimes also total GM volume] using partial correlations, and were subsequently imported into R software for the Bayesian correlation test”. This leaves an ambiguity as to whether a correction for nuisance variables was applied to both the structural data and the behavioural measure, or only the structural data. If the latter, then the correlation between the nuisance-adjusted structural data and the unadjusted behavioural measure may not be identical to the partial correlations carried out in the original studies (this is, in a sense, the complement to the distinction between partial and semipartial or part correlations, see e.g. Cohen, Cohen, West, & Aiken, 2013). If the behavioural measures are correlated with one or more of the nuisance variables, for example showing age or gender effects, then the partial correlation may be underestimated by the incomplete adjustment for the nuisance variables.

Fourth, although Figure 1 in Boekel et al. highlights the potential of Bayesian inference to combine all of the available data – original (and in some cases original in-study replication) and replication – in an updated posterior distribution for the effect, such pooling of evidence seems not to be performed outside of this figure. The figure shows that with an original $r = 0.93$ and a “failed” replication $r = 0.03$, the pooled-data posterior distribution still has appreciable mass over positive correlations, with a maximum at approximately $r = 0.4$, suggesting that several of the “failed” replications might still lead to pooled-data posterior distributions favouring an effect.

Focusing now on the VBM studies, while replications 2 & 4 (Kanai, Bahrami, Roylance, & Rees, 2012; Kanai, Dong, Bahrami, & Rees, 2011) used the same conceptual morphometric approach (apart from their use of ROIs instead of voxel-wise statistics), they used different software with different values for certain options. The original studies used the version of VBM implemented in version 8 of the statistical parametric mapping software (SPM8; Wellcome Trust Centre for Neuroimaging, UCL Institute of Neurology, <http://www.fil.ion.ucl.ac.uk>), with diffeomorphic anatomical registration through exponentiated Lie-algebra or DARTEL, however the replication attempt was carried out using VBM as implemented in FSL (the FMRIB Software Library, wherein FMRIB is the Oxford Centre for Functional MRI of the Brain; Douaud et al. (2007)). Before pointing out differences between these methods it is worth considering the basic principles underlying VBM. Automatic measurement of grey matter volumes (i.e. without laborious and potentially biased manual definition; see Ioannidis, 2011) can be done by registering individual brains to a common template; the amount of stretching or squashing needed to move voxels within a structure from its original native space into the common space provides an index of that structure’s volume (i.e. VBM). This apparent simplicity of registering two brains together

belies its complexity at both conceptual and computational levels (Crum, Griffin, Hill, & Hawkes, 2003). The matching should be similar but should not be biologically implausible. For instance, the brain of a rat could be digitally manipulated to appear visually similar to that of a human, but the resulting information would be meaningless (Rohlfing, 2012). Similarly, apparently exact matching of any two human brains is possible but normal variation in brain structure, such as differences in the number or location of sulci, could lead to abnormally large gyri, distended white matter structures or otherwise inexact homology. Current VBM analysis routines in SPM and FSL carefully balance the need to compare small scale differences in the local composition of brain tissue, while discounting large scale differences in gross anatomy and position. This is helped by registering images to a template brain, which in recent versions of VBM, such as SPM's DARTEL or FSL's FNIRT used in FSL-VBM, is created from an average image of the participants' own tissue segmentations (Ashburner & Friston, 2009). This reduces the differences in gross anatomy and so improves registration quality and hence subsequent sensitivity.

VBM analyses in SPM and FSL differ in their method for preprocessing. It has largely been assumed that findings should be robust to these approaches but a recent study demonstrates otherwise (Rajagopalan et al., 2014). In their study, Rajagopalan et al. found that a number of factors that differ between SPM and FSL can influence findings, from relatively small differences introduced by their segmentation algorithms to larger differences introduced by their registration algorithms.

To illustrate the differences in segmentation and registration, intended to complement the results of Rajagopalan, we used both SPM and FSL to generate template images from the Boekel et al. data. As can be seen in Figure 1, the two templates are broadly similar (i.e. both may be described simplistically as being "in MNI space"), but show many differences in the details. The method implemented in SPM was associated with greater grey matter probabilities in outer cortical regions. In contrast, FSL displayed greater grey matter probabilities in inner cortical regions, and numerous small differences in alignment can be observed. We then measured the voxel-by-voxel correlation in signal intensity (probabilistic tissue volume) between images processed using SPM and the same images processed using FSL (for consistency with SPM, we modified the FSL-VBM pipeline so that the modulation step preserved original volume by including the affine determinant, rather than the default non-linear-only modulation). In both cases images were smoothed using an 8 mm full-width at half-maximum Gaussian kernel. We applied a mask to ensure values were only measured in voxels with an average intensity (across both SPM and FSL) of 0.2. Across these masked voxels, the mean correlation between signal intensities was $r = 0.67$, and the median correlation was $r = 0.69$ (minimum correlation = -0.08 , maximum = 0.96). Figure 2 shows the relatively sparse regions that have a correlation of at least $r = 0.8$, which could be considered relatively lenient, given these processed images are generated from the same data. Boekel et al.'s use of ROIs provided by the original authors implicitly assumes matching image processing pipelines; it is difficult to know how much inaccuracy is introduced by the mismatch, but given the concerns about the low sample size, any further increase in noise and reduction in sensitivity is potentially problematic.

In the case of the Boekel study, the authors had originally pre-registered their design, stipulating the exact software and analysis they would use. Appropriate peer-review may have detected the potential confound in methodology earlier, and this should be considered in future pre-registration efforts. These concerns were pointed out during review of the original manuscript (by NM), however the reasonable response of Boekel and colleagues was that they had conducted the study in accordance with their pre-registered protocol and that all data would be publicly available, allowing re-analysis using SPM and any other appropriate method, as has been carried out here and elsewhere. This allowed a sensible balance between following pre-registered methods and providing the opportunity for faithful replication of the original work.

Differences between methods may have contributed to the failure to replicate in the Boekel et al. study but it is unlikely that this explains all of their findings. However, it is important to keep in mind that the reliability of many structural brain imaging studies has been established with converging validity from multiple *in vivo* imaging, post-mortem histological and computational simulation sources. True, reliable effects often require large studies or meta-analyses (see Fusar-Poli et al., 2014). Replication attempts can help greatly to establish the strength of different effects – especially, as noted earlier, where the original studies focused solely on localisation rather than non-circular quantification of detected effects. We would suggest that the ideal neuroimaging replication study would have a sample size large enough to permit an internal split-half cross-validation approach, where a conventional mass-univariate search could then be complemented with a non-circular measurement of the effect, without any confounds arising from methodological differences. Failing that, matching the original methods as closely as possible would help to narrow the range of possible interpretations of any apparently null replications. As the first major replication study of structural brain-behaviour relationships, Boekel et al. make a major step for the field, but their “failed replications” should be interpreted cautiously, rather than seen simplistically as a set of conclusively absent effects, a conclusion that some readers and some of the media might be prone to reach even when the authors of the replication have been more careful. Further discussion of the limitations, and strengths, of VBM and other structural imaging association studies can help to ensure future studies are robust and allow greater understanding of the structural basis of behaviour.

References

- Ashburner J, Friston KJ. Computing average shaped tissue probability templates. *NeuroImage*. 2009; 45(2):333–341. [PubMed: 19146961]
- Bergouignan L, Chupin M, Czechowska Y, Kinkingnehun S, Lemogne C, Le Bastard G, Lepage M, Garnero L, Colliot O, Fossati P. Can voxel based morphometry, manual segmentation and automated segmentation equally detect hippocampal volume differences in acute depression? *NeuroImage*. 2009; 45(1):29–37. [PubMed: 19071222]
- Boekel W, Wagenmakers EJ, Belay L, Verhagen J, Brown S, Forstmann BU. A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*. 2015; 66:115–133. [PubMed: 25684445]
- Bookstein FL. “Voxel-based morphometry” should not be used with imperfectly registered images. *NeuroImage*. 2001; 14(6):1454–1462. [PubMed: 11707101]

- Chambers CD, Feredoes E, Muthukumaraswamy SD, Etchells P. Instead of “playing the game” it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*. 2014; 1(1):4–17.
- Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge; 2013.
- Crum WR, Griffin LD, Hill DL, Hawkes DJ. Zen and the art of medical image registration: correspondence, homology, and quality. *NeuroImage*. 2003; 20(3):1425–1437. [PubMed: 14642457]
- den Braber A, Bohlken MM, Brouwer RM, van't Ent D, Kanai R, Kahn RS, de Geus EJ, Hulshoff Pol HE, Boomsma DI. Heritability of subcortical brain measures: a perspective for future genome-wide association studies. *NeuroImage*. 2013; 83:98–102. [PubMed: 23770413]
- DeYoung CG, Hirsh JB, Shane MS, Papademetris X, Rajeevan N, Gray JR. Testing predictions from personality neuroscience. *Brain structure and the big five*. *Psychological Science*. 2010; 21(6):820–828. [PubMed: 20435951]
- Dienes, Z. *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan; Hampshire: UK.: 2008.
- Douaud G, Smith S, Jenkinson M, Behrens T, Johansen-Berg H, Vickers J, James S, Voets N, Watkins K, Matthews PM, James A. Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*. 2007; 130(Pt 9):2375–2386. [PubMed: 17698497]
- Durston S, Fossella J, Casey B, Pol HH, Galvan A, Schnack H, Steenhuis M, Minderaa R, Buitelaar J, Kahn R. Differential effects of DRD4 and DAT1 genotype on fronto-striatal gray matter volumes in a sample of subjects with attention deficit hyperactivity disorder, their unaffected siblings, and controls. *Molecular Psychiatry*. 2005; 10(7):678–685. [PubMed: 15724142]
- Fusar-Poli P, Radua J, Frascarelli M, Mechelli A, Borgwardt S, Di Fabio F, Biondi M, Ioannidis JP, David SP. Evidence of reporting biases in voxel-based morphometry (VBM) studies of psychiatric and neurological disorders. *Human Brain Mapping*. 2014; 35(7):3052–3065. [PubMed: 24123491]
- Ioannidis JP. Excess significance bias in the literature on brain volume abnormalities. *Arch Gen Psychiatry*. 2011; 68(8):773–780. [PubMed: 21464342]
- Johnson SL, Carver CS, Joormann J. Impulsive responses to emotion as a transdiagnostic vulnerability to internalizing and externalizing symptoms. *J Affect Disord*. 2013; 150(3):872–878. [PubMed: 23726781]
- Kanai R, Bahrami B, Roylance R, Rees G. Online social network size is reflected in human brain structure. *Proc Biol Sci*. 2012; 279(1732):1327–1334. [PubMed: 22012980]
- Kanai R, Dong MY, Bahrami B, Rees G. Distractibility in daily life is reflected in the structure and function of human parietal cortex. *Journal of Neuroscience*. 2011; 31(18):6620–6626. [PubMed: 21543590]
- Kanai R, Rees G. The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews: Neuroscience*. 2011; 12(4):231–242.
- Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*. 2009; 12(5):535–540. [PubMed: 19396166]
- Kriegeskorte N, Lindquist MA, Nichols TA, Poldrack RA, Vul E. Everything you never wanted to know about circular analysis, but were afraid to ask. *J Cereb Blood Flow Metab*. 2010; 30(9):1551–1557. [PubMed: 20571517]
- Leube DT, Weis S, Freymann K, Erb M, Jessen F, Heun R, Grodd W, Kircher TT. Neural correlates of verbal episodic memory in patients with MCI and Alzheimer's disease--a VBM study. *Int J Geriatr Psychiatry*. 2008; 23(11):1114–1118. [PubMed: 18449954]
- Matsuo K, Nicoletti M, Nemoto K, Hatch JP, Peluso MA, Nery FG, Soares JC. A voxel-based morphometry study of frontal gray matter correlates of impulsivity. *Human Brain Mapping*. 2009; 30(4):1188–1195. [PubMed: 18465751]
- Posthuma D, De Geus EJ, Baaré WF, Pol HEH, Kahn RS, Boomsma DI. The association between brain volume and intelligence is of genetic origin. *Nature Neuroscience*. 2002; 5(2):83–84. [PubMed: 11818967]

- Radua J, Canales-Rodriguez EJ, Pomarol-Clotet E, Salvador R. Validity of modulation and optimal settings for advanced voxel-based morphometry. *NeuroImage*. 2014; 86:81–90. [PubMed: 23933042]
- Rajagopalan V, Yue GH, Pioro EP. Do preprocessing algorithms and statistical models influence voxel-based morphometry (VBM) results in amyotrophic lateral sclerosis patients? A systematic comparison of popular VBM analytical methods. *Journal of Magnetic Resonance Imaging*. 2014; 40(3):662–667. [PubMed: 24470336]
- Ridgway GR, Henley SM, Rohrer JD, Scahill RI, Warren JD, Fox NC. Ten simple rules for reporting voxel-based morphometry studies. *NeuroImage*. 2008; 40(4):1429–1435. [PubMed: 18314353]
- Ridgway GR, Omar R, Ourselin S, Hill DL, Warren JD, Fox NC. Issues with threshold masking in voxel-based morphometry of atrophied brains. *NeuroImage*. 2009; 44(1):99–111. [PubMed: 18848632]
- Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Transactions on Medical Imaging*. 2012; 31(2):153–163. [PubMed: 21827972]
- Shen S, Sterr A. Is DARTEL-based voxel-based morphometry affected by width of smoothing kernel and group size? A study using simulated atrophy. *Journal of Magnetic Resonance Imaging*. 2013; 37(6):1468–1475. [PubMed: 23172789]
- Tardif CL, Collins DL, Pike GB. Sensitivity of voxel-based morphometry analysis to choice of imaging protocol at 3 T. *NeuroImage*. 2009; 44(3):827–838. [PubMed: 18996205]
- Tardif CL, Collins DL, Pike GB. Regional impact of field strength on voxel-based morphometry results. *Human Brain Mapping*. 2010; 31(7):943–957. [PubMed: 19862698]
- Toga AW, Thompson PM. Genetics of brain structure and intelligence. *Annual review of Neuroscience*. 2005; 28:1–23.
- Vul E, Harris C, Winkielman P, Pashler H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on psychological science*. 2009; 4(3):274–290. [PubMed: 26158964]
- Westlye LT, Grydeland H, Walhovd KB, Fjell AM. Associations between regional cortical thickness and attentional networks as measured by the attention network test. *Cerebral Cortex*. 2011; 21(2):345–356. [PubMed: 20525771]
- Winkler AM, Kochunov P, Blangero J, Almasy L, Zilles K, Fox PT, Duggirala R, Glahn DC. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage*. 2010; 53(3):1135–1146. [PubMed: 20006715]
- Woollett K, Maguire EA. Acquiring “the Knowledge” of London’s layout drives structural brain changes. *Current Biology*. 2011; 21(24):2109–2114. [PubMed: 22169537]
- Wright IC, Sham P, Murray RM, Weinberger DR, Bullmore ET. Genetic contributions to regional variability in human brain structure: methods and preliminary results. *NeuroImage*. 2002; 17(1):256–271. [PubMed: 12482082]
- Zapolski TC, Guller L, Smith GT. Construct validation theory applied to the study of personality dysfunction. *J Pers*. 2012; 80(6):1507–1531. [PubMed: 22321263]

Figure 1.

Template images generated from (unmodulated) probabilistic segmented grey matter maps in SPM (first column) and in FSL (second column), overlaid on an average of the original images normalised using SPM's DARTEL transformations. The third column demonstrates the differences between template images, windowed between 0.25 and 0.75; in these images greater GM probabilities in SPM images are shown in yellow-red, and greater GM probabilities in FSL are shown in blue-light blue.

Figure 2. Regions showing a relatively high ($r = 0.8$ or greater) correlation between levels of signal intensity after preprocessing (smoothing, modulation, spatial normalisation, and tissue segmentation) using SPM and FSL.