Psychological Medicine (2015), **45**, 449–451. © Cambridge University Press 2014
doi:10.1017/S0033291714001354

EDITORIAL

Has analytical flexibility increased in imaging studies of bipolar disorder and major depression?

M. R. Munafò[1,2]* and M. J. Kempton[3]

[1] *UK Centre for Tobacco and Alcohol Studies and School of Experimental Psychology, University of Bristol, Bristol, UK*
[2] *MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, UK*
[3] *Department of Neuroimaging, Institute of Psychiatry, King's College London, London, UK*

There has been extensive discussion of problems of reproducibility of research. Analytical flexibility may contribute to this, by increasing the likelihood that a reported finding represents a chance result. We explored whether analytical flexibility has increased over time, using human imaging studies of bipolar disorder and major depression. Our results indicate that the number of measures collected per study has increased over time for studies of bipolar disorder, but not for studies of major depression.

Received 5 February 2014; Revised 14 May 2014; Accepted 20 May 2014; First published online 25 June 2014

Key words: Analytical flexibility, bipolar disorder, imaging, major depression.

There has been extensive discussion of problems of reproducibility of research across a range of scientific disciplines (Ioannidis, 2005). A number of factors have been identified that may contribute to this, such as data fabrication (Simonsohn, 2013), publication bias (Smulders, 2013), peer review methods (Park *et al.* 2013) and low statistical power (Button *et al.* 2013). For the most part these are not new concerns; however, one factor that may have changed over recent years is the scope for flexible data analysis, given the increasing automation of statistical analyses, and the ease with which multiple outcomes can be tested in the same dataset.

The impact of flexible analytical procedures has recently been described by Simmons *et al.* (2011), who concluded that it is 'unacceptably easy to accumulate (and report) statistically significant evidence for a false hypothesis'. This problem is not confined to behavioural experiments in psychology – Carp (2012) recently reviewed 241 functional magnetic resonance imaging (MRI) studies and showed that there were almost as many unique analytical pipelines reported as there were individual studies, with many studies not reporting critical methodological details. However, it is not clear whether analytical flexibility has increased over time.

We therefore investigated whether analytical flexibility in structural imaging studies of bipolar disorder and major depression has increased over time, using the number of measures collected as a proxy index of analytical flexibility. With more measures available, there is greater scope for conducting multiple statistical tests, and selecting those that provide the clearest results for reporting or highlighting. We also investigated whether the number of participants tested has increased over time.

Data were taken from the Bipolar Disorder Neuroimaging Database (bipolardatabase.org) (Kempton *et al.* 2008) and the Major Depression Neuroimaging Database (depressiondatabase.org) (Kempton *et al.* 2011). These online databases include peer-reviewed computerized tomography and structural MRI studies that compare patients with bipolar disorder or patients with major depression, diagnosed using standard diagnostic criteria, with a healthy control group. From studies within these two databases, the total number of participants (patients and controls) and total number of different brain measures recorded per study were extracted for the present analyses. Brain measures were defined as the measurement of a brain region (e.g. left hippocampus volume) or the measurement of a cerebral abnormality (e.g. the presence of periventricular hyperintensities).

We used linear regression to explore the relationship between year of publication, number of measures and number of participants. For studies of bipolar disorder (k=141), year of publication was not associated with number of participants [B=−0.01, 95% confidence interval (CI) −0.02 to 0.01, R^2=0.01, p=0.23] but

* Address for correspondence: M. R. Munafò, Ph.D., School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK.
(Email: marcus.munafo@bristol.ac.uk)

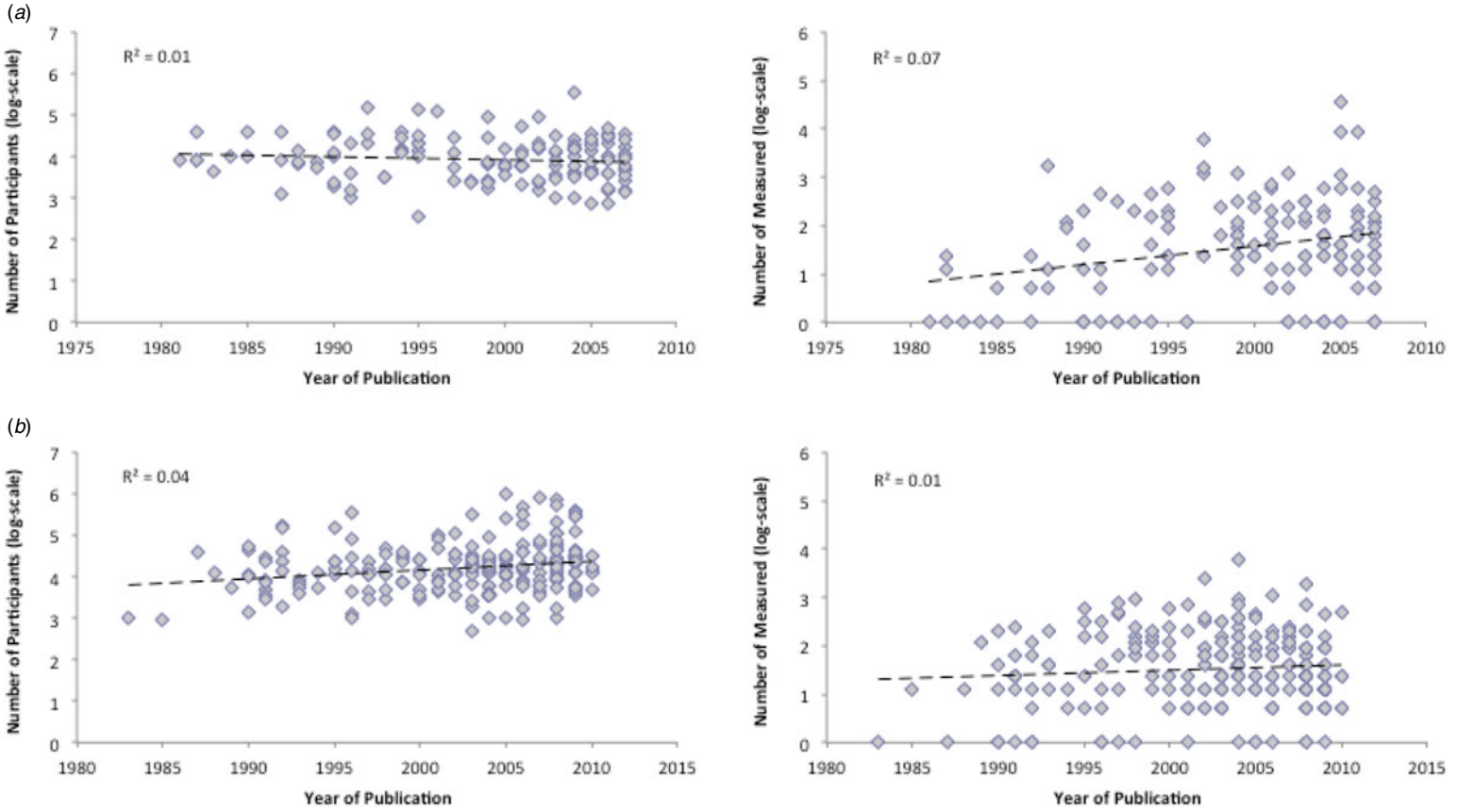

Fig. 1. Association of year of publication with number of participants and number of measures in structural imaging studies of major depression and bipolar disorder. In studies of bipolar disorder (*a*), year of publication is not associated with number of participants (R^2=0.01, p=0.23) but is associated with number of measures collected (R^2=0.07, p=0.001). However, in studies of major depression (*b*), year of publication is associated with number of participants (R^2=0.04, p=0.001) but not with number of measures collected (R^2=0.01, p=0.21).

was positively associated with number of measures (B=0.04, 95% CI 0.02–0.06, R^2=0.07, p=0.001). However, for studies of major depression (k=225), year of publication was positively associated with number of participants (B=0.02, 95% CI 0.01–0.03, R^2=0.04, p=0.001) but not with number of measures (B=0.01, 95% CI −0.01 to 0.03, R^2=0.01, p=0.21). A Z test indicated that these estimates differed, with strong evidence for number of participants (p=0.004) and weaker evidence for number of measures (p=0.080). These results are shown in Fig. 1.

Our results partly support the possibility that analytical flexibility has increased over time. Among structural imaging studies of bipolar disorder, the number of measures taken per study (assumed here to be a proxy index of analytical flexibility) has increased, while the average sample size has not. However, among structural imaging studies of major depression we observed the opposite pattern, with no increase in the number of measures taken per study but an increase in average sample size. The reasons for this discrepancy are not clear. We restricted our analysis to structural MRI region-of-interest studies because relevant analysis techniques are well established, and therefore consistent across studies. While strength of the MRI scanner and slice thickness may influence results, we previously found no evidence that these factors influenced measures of six key brain regions (Kempton *et al.* 2011). It is possible that our results represent chance findings, but the statistical evidence is sufficiently strong that this explanation is unlikely. The results also do not appear to be driven by a small number of outliers.

One possibility is that there are in fact fewer true effects in bipolar disorder compared with major depression (or the effects are considerably smaller). If it is harder to detect effects this may lead to increased pressure to collect multiple measures to increase the likelihood of finding something. The addition of future study databases recording analytical flexibility may clarify the apparent discrepancy between the major depression and bipolar disorder literatures. More generally, there is growing interest in methods to interrogate published literature. Our approach, which uses number of measures as a metric of analytical flexibility, may be useful as a scalable tool for analysing all available studies across a published literature.

Acknowledgements

M.R.M. is a member of the United Kingdom Centre for Tobacco and Alcohol Studies, a UK Clinical Research Collaboration (UKCRC) Public Health Research Centre of Excellence. Funding from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UKCRC, is gratefully acknowledged. M.J.K. is supported by a Medical Research Council Career Development Fellowship (grant no. MR/J008915).

Declaration of Interest

None.

References

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376.

Carp J (2012). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* **63**, 289–300.

Ioannidis JP (2005). Why most published research findings are false. *PLoS Medicine* **2**, e124.

Kempton MJ, Geddes JR, Ettinger U, Williams SC, Grasby PM (2008). Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Archives of General Psychiatry* **65**, 1017–1032.

Kempton MJ, Salvador Z, Munafò MR, Geddes JR, Simmons A, Frangou S, Williams SC (2011). Structural neuroimaging studies in major depressive disorder. Meta-analysis and comparison with bipolar disorder. *Archives of General Psychiatry* **68**, 675–690.

Park IU, Peacey MW, Munafò MR (2013). Modelling the effects of subjective and objective decision making in scientific peer review. *Nature* **506**, 93–96.

Simmons JP, Nelson LD, Simonsohn U (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**, 1359–1366.

Simonsohn U (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological Science* **24**, 1875–1888.

Smulders YM (2013). A two-step manuscript submission process can reduce publication bias. *Journal of Clinical Epidemiology* **66**, 946–947.