BJPsych The British Journal of Psychiatry (2010)
197, 257–258. doi: 10.1192/bjp.bp.109.069849

Editorial

How reliable are scientific studies?

Marcus R. Munafò and Jonathan Flint

Summary
There is growing concern that a substantial proportion of scientific research may in fact be false. A number of factors have been proposed as contributing to the presence of a large number of false-positive results in the literature, one of which is publication bias. We discuss empirical evidence for these factors.

Declaration of interest
Both authors hold opinions which are likely to influence their interpretation of evidence, including that presented here.

Marcus R. Munafò (pictured) is Reader in Biological Psychology at the University of Bristol, with expertise in the biological basis of behavioural traits. Jonathan Flint is Michael Davys Professor at the University of Oxford, with expertise in translational models of the genetic basis of behavioural traits.

'One of the strengths of science is that it does not require that scientists are unbiased, only that different scientists have different biases.'
David Hull, *Science as a Process*.

Scientific discovery and chance

During the Second World War, the physicist Enrico Fermi asked General Leslie Groves how many generals might be called 'great', and why. Groves replied that any general who won five major battles in a row might be called 'great', and that about 3 in every 100 would qualify. Fermi countered that if opposing forces are roughly equal, the odds are 1 in 2 that a general will win one battle, 1 in 4 that he will win two battles in a row, 1 in 8 for three battles, 1 in 16 for four battles, and 1 in 32 for five battles in a row. 'So you are right, General, about three in a hundred. Mathematical probability, not genius'. In other words, apparently striking consistency may only be the consequence of the inexorable laws of probability. In this editorial we suggest that, by the same inexorable logic, many scientific discoveries might be called 'great'.

An analogue of Fermi's 'great General' may be the 'great scientific discovery' – apparently exciting findings often subsequently fail to replicate, and may have originally occurred simply owing to chance, given the sheer amount of scientific research that is conducted. Here, we take as an example the work of researchers investigating the relationship between disease susceptibility and DNA sequence variants, using genetic association studies.

To outsiders, the odds are 1 in 20 that a correlation (in this case a genetic association) will be observed if there is in fact no association (assuming that a scientific journal accepts a *P* threshold of 0.05 as sufficient evidence for publication) and 1 in 400 that the discovery will be replicated by chance, providing a reasonable level of confidence that most replicated findings are real. But for many (if not the majority) of studies, the odds in favour of publication may be much lower for both discovery and replication. Statistical software packages enable researchers to conduct multiple statistical tests at astonishing speed, and it has become routine to do so. One recent realistic simulation study, using ten sequence variants in the widely studied gene for the catechol-*O*-methyltransferase (COMT) enzyme and a package of analyses similar to those employed in practice, reported a false-positive rate of 96.8% at the P=0.05 level of significance.[1] Furthermore, under a loose definition of replication, spurious findings 'replicated' in the majority of cases, again using random data.

Does this happen in practice? Although empirical evidence of an excess of *P*-values just below the 5% threshold indicates that researchers frequently do run multiple tests on their data,[2] we believe that false-positive findings permeate the literature for additional reasons. We have pointed out that one of the most influential and highly cited reports in behaviour genetics, in which susceptibility to depression is claimed to depend upon the presence of a particular allele of the serotonin transporter gene, is most likely due to chance.[3] Analysis of the different ways in which interactions between genetic variants and life stresses were claimed as replication showed that the nature of the interaction in the replication study was often ignored; consequently, replications were not, in the majority of cases, strict replications of the original finding.

Furthermore, low statistical power appears to be endemic in many fields. We have investigated genetic association studies,[4] neuroimaging phenotypes[5] and laboratory paradigms for assessing responsivity to environmental cues in drug users,[6] and in all cases found the average statistical power (based on the median sample size of studies in each respective meta-analysis) to be roughly between 15 and 25% (Fig. 1). If these values are representative, this means that if 90% of our hypotheses are in fact null, and we retain an alpha level of 5%, the majority of statistically significant (and therefore, presumably, published) findings will in fact be false.[7]

What undermines the reliability of studies?

Why is so much scientific research likely to be false? A number of factors are empirically known to introduce bias into the literature and contribute to the risk of false-positive results: publication bias; longer time to publish for results which do not achieve statistical significance; the trend for effect sizes to decrease with year of publication; the poor predictive value of initial reports; the *post hoc* study of further subgroups defined by gender or environmental factors; and source of funding. There is evidence that all of these frequently occur.

However, there are other sources of bias within the social fabric of science which are less well described and under-researched. For example, we used data from three meta-analytic reviews of gene–disease associations in the psychiatric genetics literature, and estimated the degree to which each individual study over- or underestimated the true effect size (from the corresponding meta-analysis). We found, perhaps paradoxically, that studies published in journals with a low impact factor are more likely to give an accurate effect size estimate than those published in

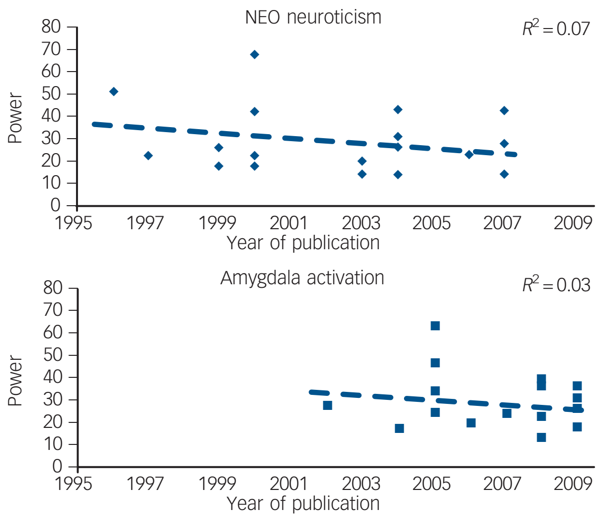

Fig. 1 Statistical power of genetic association studies of neuroticism and amygdala activation.

Statistical power of individual studies is presented against year of publication for studies of the 5-HTTLPR genetic variant and measures of both neuroticism (assessed using the NEO personality questionnaire) and amygdala activation, based on the effect sizes in the corresponding meta-analysis. In both cases, power has remained low over several years, despite growing evidence that studies are underpowered. Low power increases the proportion of false-positive to true-positive findings among those studies that achieve nominal statistical significance. Data adapted and updated from Munafò *et al.*[4,5]

journals with a high impact factor.[8] We also found evidence that the location where a study is conducted is associated with the degree to which it represents an overestimate of the true effect size, with studies conducted in North America overestimating the likely true effect size by around 10% compared with those conducted in Europe and elsewhere.[9]

It is likely that subtle factors serve to influence the reporting of scientific studies,[10] and in 'hot' scientific fields where there is substantial flexibility in study design there is perhaps greater scope for these factors to play a role.[7] Much of the evidence we have presented comes from molecular genetic observational studies, but there is no reason to suspect that this field is a particular culprit. Rather, the large numbers of relatively comparable studies allow the investigation of extra-scientific factors to a greater degree than in other fields, where attempted replication is less common. This indifference to replication in some fields is itself a problem.

What can we do?

Can we do anything to improve this situation? Reviewers, journal editors and science policy markers could enforce higher standards, taking the clinical trials literature as an example of good practice. For example, pre-publication of study protocols, to discourage deviation from planned analyses, as well as triple-blind data collection and analysis, all serve to minimise unnecessary statistical testing, discourage 'data mining', and facilitate transparent reporting, while the routine use of power analysis to determine sample size reduces the ratio of false-positive to true-positive findings. There is perhaps a need for evidence-based science, as well as evidence-based medicine.

In the meantime, readers of scientific journals should perhaps only believe large studies which report on findings in a mature literature (as opposed to early findings in a new field), place less emphasis on nominal statistical significance and focus instead on effect sizes and confidence intervals, and are published in journals with a low impact factor. Many of the problems highlighted above are increasingly recognised within the psychiatric genetics literature, reflected in the use of much larger samples to achieve sufficient statistical power, a requirement for robust replication before findings are regarded as even tentatively established, and a wider discussion of statistical issues and in particular Bayesian approaches.[11] This is a positive move, and indicates that science has the potential to correct itself by identifying these problems, so that we can learn from these and subsequently improve our methods. More generally, we should be aware that biases can take many forms, beyond the usual suspects of financial vested interests and source of research funding, and are likely to operate across all domains of scientific enquiry. We should accept that definitive answers require definitive (which generally means large, but also high-quality) studies, and perhaps focus on doing less science, but doing it better.

Marcus R. Munafò, PhD, Department of Experimental Psychology, University of Bristol; **Jonathan Flint**, FRCPsych, Wellcome Trust Centre for Human Genetics, University of Oxford, UK.

Correspondence: Marcus R. Munafò, Department of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK. Email: marcus.munafo@bristol.ac.uk

First received 1 Jul 2009, final revision 10 Nov 2009, accepted 2 Dec 2009

Funding

M.R.M. is supported by the Higher Education Funding Council for England (HEFCE). J.F. is supported by the Wellcome Trust.

Acknowledgements

Published scholarly articles were used as sources of information for the article. M.R.M. is guarantor for the article.

References

1 Sullivan PF. Spurious genetic associations. *Biol Psychiatry* 2007; **61**: 1121–6.

2 Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials* 2007; **4**: 245–53.

3 Munafò MR, Durrant C, Lewis G, Flint J. Gene x environment interactions at the serotonin transporter locus. *Biol Psychiatry* 2009; **65**: 211–9.

4 Munafò MR, Freimer NB, Ng W, Ophoff R, Veijola J, Miettunen J, et al. 5-HTTLPR genotype and anxiety-related personality traits: a meta-analysis and new data. *Am J Med Genet B Neuropsychiatr Genet* 2009; **150B**: 271–81.

5 Munafò MR, Brown SM, Hariri AR. Serotonin transporter (5-HTTLPR) genotype and amygdala activation: a meta-analysis. *Biol Psychiatry* 2008; **63**: 852–7.

6 Field M, Munafò MR, Franken IH. A meta-analytic investigation of the relationship between attentional bias and subjective craving in substance abuse. *Psychol Bull* 2009; **135**: 589–607.

7 Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005; **2**: e124.

8 Munafò MR, Stothart G, Flint J. Bias in genetic association studies and impact factor. *Mol Psychiatry* 2009; **14**: 119–20.

9 Munafò MR, Attwood AS, Flint J. Bias in genetic association studies: effects of research location and resources. *Psychol Med* 2008; **38**: 1213–4.

10 Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. *Nature* 2005; **435**: 737–8.

11 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–78.

https://doi.org/10.1192/bjp.bp.109.069849 Published online by Cambridge University Press