How reliable are scientific studies?

Marcus R. Munafo 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.

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.

Furthermore, under a loose definition of replication, spurious findings 'replicated' in the majority of cases, again using random data. 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

Declaration of interest
Both authors hold opinions which are likely to influence their interpretation of evidence, including that presented here.
journals with a high impact factor. 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.

It is likely that subtle factors serve to influence the reporting of scientific studies, 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. 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. Where 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 ‘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

References

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.