PAUL HACKETT/IN PICTURES/CORBIS

The purported decline in bee numbers raises questions about evidence quality.

A standard for policy-relevant science

Ian Boyd calls for an auditing process to help policy-makers to navigate research bias.

The increasing concern about unreliability in scientific literature[1,2] is a problem for people like me — I am the science adviser to DEFRA, the UK government department for environment, food and rural affairs. To counsel politicians, I must recognize systematic bias in research. Bias is cryptic enough in individual studies, let alone in whole bodies of literature that contain important inaccuracies[2,3].

It worries me that because of bias, some parts of the published scientific literature, such as studies on the safety of genetically modified (GM) organisms and pesticides, or trends in biodiversity measurements, might have only limited use in policy-making.

To mitigate this problem, policy-makers should consider holding published scientific evidence to an audited standard that can be replicated and is robust to variations in assessor competence. A weighting factor, or 'kite mark', applied to journals or individual articles, could help policy-makers to assess the robustness of studies for use in particular applications. Similar methods established by non-profit standards associations are used in research to certify laboratory practice and in engineering to certify building standards.

The quality of research results fluctuates because of varying tractability in the problems being probed[1,4]. For example, it is easy to judge the efficacy of an experiment to engineer a tomato to produce the pigment anthocyanin[5], because if it succeeds, that tomato is the colour of a ripe plum. It is much harder to judge the reliability of a study investigating whether a GM crop is toxic to animals[6]. The latter situation is much more susceptible to inaccuracy and interpretation.

These problems are amplified in complex issues such as the environmental effects of GM organisms or chemical pollutants, including pesticides and endocrine disrupters. In these cases, experimentation is needed at scales large enough to provide statistical power in the presence of high background noise. The problem is amplified further when statistical inference is used.

SCOPING THE PROBLEM

Systematic bias across whole fields of science is even more cryptic and therefore more problematic. It could stem from the combined effects of how science is commissioned, conducted, reported and used, and also from how scientists themselves are incentivized to conduct certain research[7]. Such bias results from actively searching for a particular outcome, rather than performing balanced hypothesis testing. For example, in 2006, researchers in the United Kingdom and in the Netherlands found that the number of insect pollinators might have declined[8]. A consequent call for proposals (see go.nature.com/audhny) contained the underlying assumption that there was a decline, rather than conveying a need to establish whether current information about declines was robust.

Another problem is the tendency to treat different studies as statistically independent, even when they have emerged from connected commissioning processes and could therefore amount to multiple testing of the same hypothesis, meaning that every extra study must overcome an increasingly rigorous statistical hurdle to demonstrate efficacy. In combination, these kinds of bias can make individual or groups of studies that report certain effects seem more important than they really are. I suspect that these effects could be a factor in the continuing controversies

surrounding genetic modification and the failure of the EU regulatory system to process applications to license new GM products.

A common reaction to such controversy is to commission subject reviews or meta-analyses[1] that assess the weight of evidence for certain effects across many individual studies. Ideally, reviewers would use processes similar to those deployed in the Cochrane Reviews that inform decision-making in health care[9].

But reviews also contain pitfalls. First, they risk amplifying rather than eliminating systematic bias — which could be more common in some subjects than others. Second, they can be affected by the increasing tendency not to publish 'negative' results[4]. Meta-analyses can compound the prevalence of false positives in the literature, and can be blind to unreported true negatives. We need rules for how to deal with these issues when compiling literature reviews for policy-relevant research.

SEAL OF APPROVAL

Strict procedures govern experimental design and the evidence standards for trials that are used to determine the efficacy and safety of GM organisms, pesticides or drug therapies. But once products are licensed for use, they are often subject to less formal investigations. The same relaxation of rules applies to testing the efficacy of policy interventions. Ad hoc studies, with all the problems outlined above, can then carry disproportionate political weight when their results question the operational integrity of a licensed product, or the effectiveness of a policy[10]. Quality-control criteria are needed for these studies that are outside a regulatory framework.

We need an international audited standard that grades studies, or perhaps journals. It would evaluate how research was commissioned, designed, conducted and reported. This audit procedure would assess many of the fundamental components of scientific studies, such as appropriate statistical power; precision and accuracy of measurements; and validation data for assays and models. It would also consider conflicts of interest, actual or implied, and more challenging issues about the extent to which the conclusions follow from the data. Any research paper or journal that does not present all the information needed for audit would automatically attract a low grade.

"What I propose augments rather than replaces peer review."

Such a system would provide policy officials and others with a reliable way of assessing evidence quality, and it would drive up standards in scientific research to reverse the worrying trends that suggest underlying bias[1–4,7].

Critics will counter that my proposed certification standard would be subjective and would shift the job of assessing quality away from expert peer reviewers. But in its current form, peer review fails to set a consistent standard. What I propose augments rather than replaces peer review, and assessment could be carried out on behalf of authors, journals or users of information through the use of third-party certified auditors.

I do not underestimate the challenge of establishing such a system, but it would bring standards to scientific publishing that are common practice in other disciplines. Ultimately, this will increase the rigour and transparency around the scientific literature that is used in policy decisions. ■

Ian Boyd *is chief scientific adviser at the UK Department of Environment, Food and Rural Affairs in London. He is also professor in biology at the University of St Andrews, UK. e-mail: ian.boyd@defra.gsi.gov.uk*

1. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. *Br. Med. J.* **315,** 629–634 (1997).
2. Begley, C. G. *Nature* **497,** 433–434 (2013).
3. Fanelli, D. *PLoS ONE* **4,** e5738 (2009).
4. Fanelli, D. *Scientometrics* **90,** 891–904 (2012).
5. Zhang, Y. *et al. Curr. Biol.* **23,** 1094–1100 (2013).
6. Doull, J. *et al. Food Chem. Toxicol.* **45,** 2073–2085 (2007).
7. Fanelli, D. *PLoS ONE* **5,** e10271 (2010).
8. Biesmeijer, J. C. *et al. Science* **313,** 351–354 (2006).
9. Jadad, A. R. *et al. J. Am. Med. Assoc.* **280,** 278–280 (1998).
10. Stokstad, E. *Science* **340,** 674–676 (2013).

Bring on the evidence

It is time to probe whether the trend for patient and public involvement in medical research is beneficial, say **Sophie Petit-Zeman** and **Louise Locock**.

Involving patients and the public as partners in medical research — from deciding what to study to influencing how results are used — is an emerging force. For some, the approach is based on common sense and justice[1]. Others, such as the chief medical officer for England, Sally Davies, feel that the advice of patients and the public "invariably makes studies more effective, more credible and often more cost efficient"[2].

The Seventh Framework Programme (FP7), the European Union's current research-funding instrument, stresses[3] the importance of patient and public involvement, known as PPI. And the Patient-Centered Outcomes Research Institute in Washington DC has allocated US$68 million to a research network predicated on the principle that "the interests of patients will be central to decision-making"(see go.nature.com/mdhy6i).

PPI is a prerequisite for much UK government research funding and it is spreading among funders, health-care organizations and charities[4]. The James Lind Alliance (JLA), with which one of us (S.P-Z.) has worked since its inception in 2004, enables patients, carers and clinicians to agree on what research matters most. It explicitly excludes the pharmaceutical industry and pure researchers. After a decade of arms-length government support, the JLA is now part of the National Institute for Health Research (NIHR) based in Southampton, UK, and JLA partnerships are complete or underway for 25 medical conditions (see go.nature.com/twhvxz). For example, the NIHR Oxford Biomedical Research Centre is running partnerships in spinal-cord injury and joint-replacement surgery, and it is the first major research institution to be appointing staff to use the JLA method 'in house', closing the loop between what matters to patients and what is researched in their name.

This international growth of PPI is rightly paralleled by unease at the paucity of evidence for its impact. And the evidence there is, including the findings that PPI improves recruitment to studies and changes what is researched[2,5], is weak. As Simon Denegri, the United Kingdom's first national director for public participation and engagement in research, put it: "The evidence-base for PPI's impact is meagre, patchy and largely observational."

SELF-EXAMINATION

Those of us working in PPI must robustly examine our own practices with a common set of tools. Otherwise, we will struggle to answer PPI sceptics, such as one researcher who asked: "Why should patients have useful opinions about what directions research should take?"[6].