COMMENT

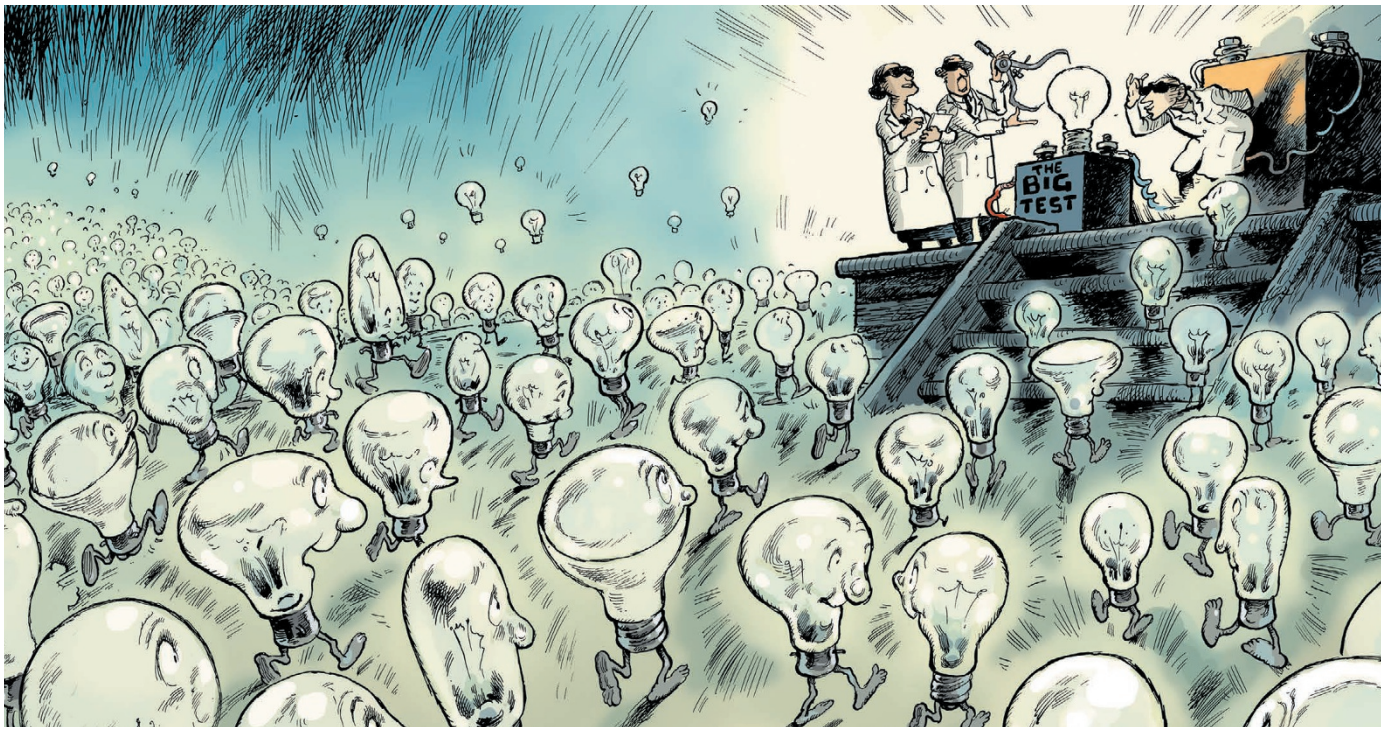
SUSTAINABILITY When the wells of the world run dry, what then? **p.412**

POPULATION The best way to reduce abortions is to invest in family planning **p.414**

PUBLISHING Pay reviewers in coupons for open-access fees **p.414**

OBITUARY Statistician Stephen E. Fienberg, remembered **p.415**

ILLUSTRATION BY DAVID PARKINS



No publication without confirmation

Jeffrey S. Mogil and **Malcolm R. Macleod** propose a new kind of paper that combines the flexibility of basic research with the rigour of clinical trials.

Concern over the reliability of published biomedical results grows unabated. Frustration with this 'reproducibility crisis' is felt by everyone pursuing new disease treatments: from clinicians and would-be drug developers who want solid foundations for the preclinical research they build on, to basic scientists who are forced to devote more time and resources to newly imposed requirements for rigour, reporting and statistics. Tightening rigour across all experiments will decrease the number of false positive findings, but comes with the risk of reducing experimental efficiency and creativity.

Bolder ideas are needed. What we propose here is a compromise between the need to trust conclusions in published papers and the freedom for basic scientists to explore and innovate[1]. Our proposal is a new type of paper for animal studies of disease therapies or preventions: one that incorporates an independent, statistically rigorous confirmation of a researcher's central hypothesis. We call this large confirmatory study a preclinical trial. These would be more formal and rigorous than the typical preclinical testing conducted in academic labs, and would adopt many practices of a clinical trial.

We believe that this requirement would push researchers to be more sceptical of their own work. Instead of striving to convince reviewers and editors to publish a paper in prestigious outlets, they would be questioning whether their hypotheses could stand up in a large, confirmatory animal study. Such a trial would allow much more flexibility in earlier hypothesis-generating experiments, which would be published in the same paper as the confirmatory study. If the idea catches on, there will be fewer high-profile papers hailing new therapeutic strategies, but much more confidence in their conclusions.

The confirmatory study would have ▶

three features. First, it would adhere to the highest levels of rigour in design (such as blinding and randomization), analysis and reporting. Second, it would be held to a higher threshold of statistical significance, such as using P values of $P < 0.01$ instead of the currently standard $P < 0.05$. Third, it would be performed by an independent laboratory or consortium. This exceeds the requirements currently proposed by various checklists and funders, but would apply only to the final, crucial confirmatory experiment.

Unlike clinical studies, most preclinical research papers describe a long chain of experiments, all incrementally building support for the same hypothesis. Such papers often include more than a dozen separate *in vitro* and animal experiments, with each one required to reach statistical significance. We argue that, as long as there is a final, impeccable study that confirms the hypothesis, the earlier experiments in this chain do not need to be held to the same rigid statistical standard.

> ***"This would represent a big shift in how scientists produce papers."***

This would represent a big shift in how scientists produce papers, but we think that the integrity of biomedical research could benefit from such radical thinking.

FINAL CONFIRMATION

For hypotheses with clear clinical implications, the logical confirmatory experiment almost always involves animal studies, in which the effect of a treatment strategy or a genetic mutation is assessed in mice or rats. The execution of these studies is often poor[2]. For example, behavioural testing — such as gauging the extent of pain or paralysis — falls outside the core competency of most molecular biology labs. Large variability or questionable baseline measures cloud results[3]. In addition, most studies conducted today have low statistical power[4] and a high risk of bias[5]. Many journals, including this one, have promoted guidelines such as those framed by the ARRIVE initiative[6]. The impact of these publishing policies is being investigated[7] but is not yet clear.

Under our proposal, a protocol for the confirmatory study would be set out in advance, specifying the hypothesis, the key outcome measures and the plan for statistical analysis. Enough animals should be studied so that a positive statistical test means that the hypothesis is very likely to be correct (see 'The maths of predictive value'). Sample sizes for this crucial experiment would need to go up; we estimate around sixfold. Overall, however, the subsequent savings in both animals and money are likely to be substantial; fewer people would waste resources following up on weak papers. This would get new drugs to market more quickly.

GETTING IT DONE

Who will conduct these hypothesis-testing experiments, and why would they want to? Preclinical trials should be run by researchers with strong expertise in the relevant animal models, and we believe that some will decide to specialize in performing confirmatory experiments for colleagues. Another option would be to establish dedicated animal-testing facilities, analogous to genomics and bioinformatics core facilities. These provide high-quality services and have become a crucial part of the scientific enterprise. Additionally, consortia might be set up to conduct such studies, and to develop and deepen the methodologies used in them.

Specialized confirmatory labs would increase the quality of animal studies, and free the labs that did the initial experiments to focus on their core expertise. We think that government funders and industry partners, which have spent billions of dollars on disappointing clinical trials, would be prepared to shift resources to support such an improved system, perhaps by offering dedicated grants. Confirmatory labs would be less dependent on positive results than the original researchers, a situation that should promote the publication of null and negative results. They would be rewarded by authorship on published papers, service fees, or both. They would also be more motivated to build a reputation for quality and competence than to achieve a particular finding.

For findings with immediate clinical applications (that is, a potential treatment that might go into human testing), we propose an extra 'generalizability study' to follow the confirmatory phase (see 'Publication with confirmation'). This would be designed to assess how widely applicable the treatment might be, and to boost confidence that it will work across a range of situations. One strategy is to repeat the confirmatory study across multiple sites, with built-in biological variability. By broadening the circumstances in which the hypothesis is tested (animal age, strain, sex, health, co-morbidity, precise assay used, drug administration, timing of outcome assessments), such studies are more likely to provide clinically useful information and to survive replication attempts.

Generalizability studies would probably be beyond an individual lab's capabilities and require multicentre consortia, but principles and tools to support them are already in place. The Multi-PART consortium (www.dcn.ed.ac.uk/multipart) has established a web-based system that allows the design, execution and assessment of studies across an unlimited number of centres. Its plans include multicentre testing of interventions that increase oxygen delivery to brain regions affected by stroke. In collaboration with the International League Against Epilepsy, it also plans to test potential new epilepsy drugs.

ENJOY THE EXPERIMENT

With a system in place for rigorous hypothesis testing, other formalities become less necessary. Any experiment in the exploratory stage could be performed without formal statistical hypothesis testing. No P-value thresholds would need to be reached; results sections might display only a central estimate, such as mean or median, and a measure of the spread of the data or, ideally, the individual data points themselves. This is in line with recommendations that the American Statistical Association made last year (see go.nature.com/2kbqkxu) that P values alone are not good measures of evidence for a hypothesis. Sample sizes should be large enough to give investigators confidence in the direction of effect, and small enough to save time and money. Complete reporting and attention to confounding variables are still essential; researchers should not exclude animals from results without mentioning them, and they should avoid methods that would introduce bias or batch effects.

PUBLICATION WITH CONFIRMATION

Our proposed paper would be accepted by journals only if it included a 'preclinical trial' following best clinical-research practices. For therapies that might later be tested in humans, all three study types are recommended.

	Exploratory studies	Confirmatory study ('preclinical trial')	Generalizability study
Who	Original researchers	Separate team, core facility or consortium	Multicentre consortium
Why	To generate hypotheses	To test hypotheses	To test broader application of hypotheses
Aims	To maximize efficiency and exploration	To avoid false positive findings	To judge readiness for clinical translation
Features	High flexibility, no mandatory statistics	High rigour and a predefined statistical analysis plan ($P<0.01$)	High rigour, built-in variability in animal subjects and assays
Publishing venue	Preprint server or informal means. Formal publication requires confirmatory study	New category of journal article, recognized as having high impact	New category of journal article, recognized as having exceptionally high impact

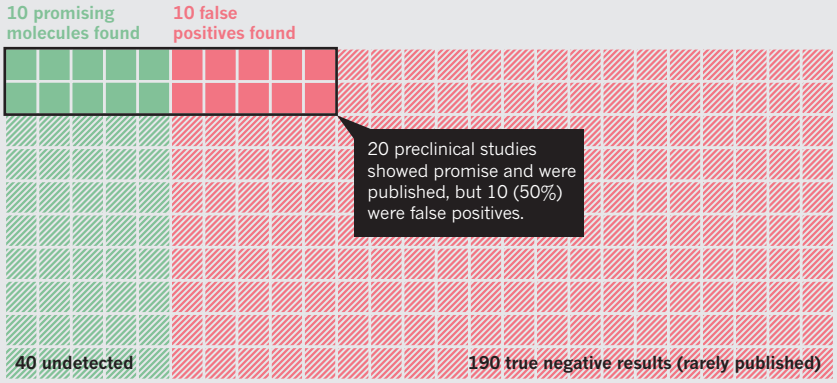
SIGNIFICANT SAMPLES

The maths of predictive value

How likely is it that a hypothesis is correct? This is best answered by positive predictive value (PPV) — not by *P* values, as is commonly thought. The PPV reflects the probability that a positive result is truly positive. It is determined by *P* values (calculated after results are collected) and statistical power (which should be calculated before a study begins). Statistical power describes the chance a study can detect some predetermined (and presumably meaningful) effect size, such as the difference between a treatment and control group.

In this example, it is assumed that 250 potential therapies went through preclinical testing. (Results later showed that 50 work (green) and 200 do not (red)). Ratios change depending on the fraction of promising molecules that actually work.

STATUS QUO: Most studies have a statistical power of only 20% and a *P* value of 0.05, meaning many more false findings (PPV of 50%). This reflects a sample size of about 10 mice per study.



PROPOSED STANDARDS: To achieve a PPV of 95%, study results would need a *P* value of 0.01 and a large enough sample size to reach 80% statistical power (typically >75 mice per study).

Wouldn't this lowered bar increase the number of false positives? We think not. Because investigators would be required to 'put up or shut up' and formally submit to a preclinical trial, they would be more comprehensive and careful with their exploratory work. They would have an incentive to do the experiments that might disprove their hypothesis at an early stage. Conversely, researchers would not feel obliged to perform experiments that they consider uninformative, as is too often the case today. Even more importantly, they would not need to increase sample size until each and every *P* value dropped below 0.05. The efficiencies gained by this change should more than overcome the resources needed to conduct a preclinical trial.

Importantly, our proposal would preserve the fun of doing exploratory science. In this new system, the costs of poor science (for example, being seduced by a rogue finding or being too cavalier in experimental design) are borne by the initial researchers. If they cut corners, cherry-pick data or eschew blinding in their experiments, they harm their chances of their hypothesis surviving the rigorous testing proposed in a preclinical trial. We predict that data fraud would decrease as well, because the need for every experiment to reach an arbitrary statistical threshold would be rendered moot. Reviewers would focus on statistics in the confirmatory study. For graduate students and postdocs, coveted publications would depend less on particular results in early experiments, and more on the strength of their overall hypotheses. Eventually, the incentive system would subtly shift to reward greater confidence and caution in conclusions: researchers would be rewarded more for the marathon than for the sprint.

This system would slow the rate of publications, but not the pace of discovery. Scientific priority could be established by the date on which an experimental plan was agreed (essentially 'registered') between the original researchers and those performing the confirmatory study. Furthermore, if published studies are more reliable and public confidence in science is boosted, a somewhat slower publication process seems acceptable. We trust that reviewers and tenure committees will find appropriate ways to credit papers that include confirmation.

WHAT NEXT?

This proposal does not fix everything that is currently broken in translational medicine, including false conclusions drawn from inappropriate animal models, unappreciated variables (such as animal microbiomes or the sex of experimenters) and publication bias. But we believe it is worth a try.

It is not practical to expect the community to change direction in step and as one. Four things could help. Journals should make space for papers that include confirmatory experiments along with exploratory work. (They could eventually prioritize them or even make confirmatory experiments a requirement.) Tenure and faculty-assessment committees should find ways to credit such work. Funders could develop schemes to pilot this approach, and those who run clinical trials should demand greater confidence in the premise underlying human studies. With even some of these incentives in place, scientists will lead the charge. ■

Jeffrey S. Mogil *is a basic neuroscientist at McGill University in Montreal, Canada.* **Malcolm R. Macleod** *is a clinical neuroscientist at the University of Edinburgh, UK.*
e-mails: jeffrey.mogil@mcgill.ca; malcolm.macleod@ed.ac.uk

1. Kimmelman, J., Mogil, J. S. & Dirnagl, U. *PLoS Biol.* **12,** e1001863 (2014).
2. Macleod, M. R. *et al. PLoS Biol.* **13,** e1002273 (2015).
3. Perrin, S. *Nature* **507,** 423–425 (2014).
4. Button, K. S. *et al. Nature Rev. Neurosci.* **14,** 365–376 (2013).
5. Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. *PLoS Biol.* **13,** e1002106 (2015).
6. Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M. & Altman, D. G. *PLoS Biol.* **8,** e1000412 (2010).
7. Cramond, F. *et al. Scientometrics* **108,** 315–328 (2016).