economically attractive. And rules should be altered to limit the amount of interest payments universities can include when calculating indirect rates, and the amount of faculty members' salaries that can be charged to grants, thereby dulling the incentive to hire people for soft-money positions. Shifting evaluations from projects to people, and de-emphasizing the importance of metrics in hiring and promotion, could encourage scientists to work on riskier projects⁶.

Many of the problems now faced by science accelerated when biomedical funding started to increase steeply. For instance, the doubling of the NIH budget from 1998 to 2003 triggered universities to hire more people and build more buildings, while scientists increased the number of grants they submitted and the size of their labs (see 'Biology's research footprint'). Now, this biomedical machine needs increasing amounts of money to sustain itself — larger labs need more grants, which leads to lower success rates, with calls for more funding.

Biomedical research has done much to contribute to increased life expectancy. But it seems likely that diminishing returns have set in. New drugs are slower in coming to market and there was a less than stellar increase in US publications associated with the NIH doubling⁷. Moreover, many of the breakthroughs that have contributed to better health outcomes have come from other fields of science — such as the laser and magnetic resonance imaging. Funds for the physical sciences in the United States (in terms of the percentage of federal research funding) are close to a 35-year low. Perhaps it is time for deans in the biomedical sciences to rent some of that excess space to their colleagues in chemistry and physics. ■

Paula Stephan is professor of economics at Georgia State University in Atlanta, Georgia 30303, USA, and a research associate at the National Bureau of Economic Research in Cambridge, Massachusetts. She is author of *How Economics Shapes Science* (Harvard Univ. Press, 2012).
e-mail: pstephan@gsu.edu

1. Franzoni, C., Scellato, G. & Stephan, P. *Science* **333**, 702–703 (2011).
2. Stephan, P. *How Economics Shapes Science* (Harvard University Press, 2012).
3. Current Population Survey, Table P-28 (United States Census Bureau, 2010); available at <http://go.nature.com/neix5i>
4. Mallon, W. T. & Korn, D. *Science* **303**, 476–477 (2004).
5. Berg, J. 'Another Look at Measuring the Scientific Output and Impact of NIGMS Grants' (22 November 2010); available at <http://go.nature.com/dae21z>
6. Azoulay, P., Graff Zivin, J. S. & Manso, G. *Incentives and Creativity: Evidence from the Academic Life Sciences* (2010); available at <http://go.nature.com/biey9d>
7. Sacks, F. *The Scientist* (11 September 2007); available at <http://go.nature.com/bpa3xr>

Turn the scientific method on ourselves

How can we know whether funding models for research work? By relentlessly testing them using randomized controlled trials, says **Pierre Azoulay**.

In times of tight budget constraints, scientists' wranglings about the real and perceived sins of public funding agencies become particularly acute. Complaints usually lead to the creation of a panel of respected, thoughtful and well-meaning scientists who come up with a plan of reform based on their intuition and experience. Funding agencies, who are genuinely concerned about improving the productivity of the scientific enterprise, often adopt these recommendations, at least in part. In one example of this process, the US National Institutes of Health (NIH) in Bethesda, Maryland, has created a large array of funding mechanisms,

each one targeted to a particular problem — including the K99/R00 or 'kangaroo' grants, which pair postdoctoral scientists with mentors to help them to prepare for tenure-track faculty positions and funding independence. Not only is this range of mechanisms confusing and costly to administer, but the effectiveness of such reforms is never seriously evaluated.

It is time to turn the scientific method on ourselves. In our attempts to reform the institutions of science, we should adhere to the same empirical standards that we insist on when evaluating research results. We already know how: by subjecting proposed reforms to a prospective, randomized ▶

ILLUSTRATION BY VIKTOR KOEN

► controlled experiment. Retrospective analyses using selected samples are often little more than veiled attempts to justify past choices.

What could such a formal experiment look like? Let me give an example. It is well documented that the past 30 years have seen a marked increase in the age at which academic scientists achieve funding independence¹. One way to ensure the continued injection of talent into these ranks would be to evaluate first-time applicants separately from a larger pool, and dedicate to them a predetermined share of the available funding. An alternative would be to keep the current system in place, but to award their proposals 'bonus points'. Which reform should we adopt? And what if the 'greying' of the scientific workforce does not stem from institutional failure, but rather reflects the influence of an ever-expanding burden of knowledge, whereby scientists must spend more time in training before they can become productive²?

To test these questions empirically, for example within the NIH (the agency I have studied most closely), we could choose a random subset of funding panels to implement the first method and a second subset to implement the other. A third subset, in which funding panels proceed with business as usual, would serve as the control group in the experiment. Ideally, the study would be designed to avoid 'panel shopping' by applicants; the 100,000 or so R01 grant proposals reviewed each year by the 183 NIH funding panels are more than enough to craft a statistical protocol with adequate power.

These experiments could exist outside government agencies, too. When philanthropic organizations develop new models to fund research, they should formally investigate how their approach compares with the dominant model, which averages experts' scores to determine the funding-priority ranking for particular projects. Some emerging models, for example, give higher ranking to projects that elicit enthusiasm and controversy than to projects that generate more consensus but only tepid support across reviewers. It might be that funding a project on the basis of reviewer sentiment is more likely to result in the selection of truly innovative, field-changing projects, but how will we know for sure? A serious evaluation of this question would compare the two systems by randomizing proposals to one of these two ranking approaches, and then examining which portfolio of projects is most successful.

When I suggest these experiments,

"Experimenting on ourselves may well lay bare some shortcomings of the scientific community."

I encounter a lot of resistance. Wouldn't this be gambling with scientists' careers? How can we measure success — by counting publications and citations, looking at the students trained as a by-product of these grants, or using other metrics? Won't this work shift scarce funding away from actual scientific investigations?

These criticisms are without merit. The current system already gambles with scientific careers, just in a haphazard way. Scientists often disagree on how to measure success, and the choice of metric, as well as the period necessary for a careful assessment, will always be context-dependent and controversial. With the good will of administrators in public agencies or private foundations, these experiments could be rolled out with minimal disruption for about the cost of a R01 grant (typically US\$250,000 per year for 5 years). If the scientific community could test even a small number of hypotheses in this way, the system-wide benefits would dwarf this modest investment.

I am well aware that this vision will sound utopian to some. Sceptics abounded when my colleagues at the Massachusetts Institute of Technology in Cambridge founded the Jameel Poverty Action Lab and subjected development-assistance methods to randomized, controlled trials to see which worked best. But as a result of this work, policy-makers now know that it is better to give out free mosquito nets to prevent malaria than to charge even a low price for their purchase — one example among many^{3,4}.

We inherited the current institutions of science from the period just after the Second World War. It would be a fortuitous coincidence if the systems that served us so well in the twentieth century were equally adapted to twenty-first-century needs. Experimenting on ourselves may well lay bare some shortcomings of the scientific community and expose us to criticisms from politicians, who are always looking for excuses to cut science funding. But the only alternative to such controlled experimentation is the gradual stultification of our most cherished scientific institutions. ■

Pierre Azoulay is an associate professor in the Massachusetts Institute of Technology Sloan School of Management, and a research associate at the National Bureau of Economic Research, Cambridge, Massachusetts, USA. e-mail: pazoulay@mit.edu

1. Rockey, S. Age Distribution of NIH Principal Investigators and Medical School Faculty (13 February 2012); available at <http://go.nature.com/2uzqca>
2. Jones, B. F. *Rev. Econ. Stud.* **76**, 283–317 (2009).
3. Cohen, J. & Dupas, P. Q. *J. Econ.* **125**, 1–45 (2010).
4. Banerjee, A. V. & Duflo, E. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty* (PublicAffairs, 2011).