CREDIT: ERIC PALMA

Peering Into Peer Review

Why don't proposals given better scores by the National Institutes of Health lead to more important research outcomes?

Michael Lauer's job at the National Institutes of Health (NIH) is to fund the best cardiology research and to disseminate the results rapidly to other scientists, physicians, and the public. But NIH's peer-review system, which relies on an army of unpaid volunteer scientists to prioritize grant proposals, may be making it harder to achieve that goal. Two recent studies by Lauer, who heads the Division of Cardiovascular Sciences at NIH's National Heart, Lung, and Blood Institute (NHLBI) in Bethesda, Maryland, raise some disturbing questions about a system used to distribute billions of dollars of federal funds each year.

Lauer recently analyzed the citation record of papers generated by nearly 1500 grants awarded by NHLBI to individual investigators between 2001 and 2008. He was shocked by the results, which appeared online last month in *Circulation Research*: The funded projects with the poorest priority scores from reviewers garnered just as many citations and publications as those with the best scores. That was the case even though low-scoring researchers had been given less money than their top-rated peers.

"Peer review should be able to tell us what research projects will have the biggest impacts," Lauer contends. "In fact, we explicitly tell scientists it's one of the main criteria for review. But what we found is quite remarkable. Peer review is not predicting outcomes at all. And that's quite disconcerting."

Two months earlier, Lauer and his NHLBI colleagues had published a study of 224 NHLBI-funded clinical trials that produced a similar bottom line, using a different marker of importance: how quickly the studies were published. Lauer believes that the two papers strongly suggest that reviewers did not do very well in separating the wheat from the chaff on $2 billion worth of NHLBI research.

Lauer says he's presented his work to NIH Director Francis Collins and other senior officials and that "not one of them pointed out any flaws or thought we had come up with some sort of erroneous finding." But that doesn't mean they agree with his conclusion.

Richard Nakamura, who oversees NIH's peer-review apparatus as head of the Center for Scientific Review (CSR), is skeptical of the use of after-the-fact yardsticks such as citations and time to publication to gauge impact. "CSR's focus is much more on what good scientists think will have high impact as opposed to what bibliometric measures might suggest will have high impact," he says.

Several social scientists who have thought about ways to measure the impact of peer review hail Lauer's willingness to put the current system under a microscope. Such studies could help NIH and other U.S. research agencies do a better job of allocating scarce resources, says economist Adam Jaffe, who directs Motu Economic and Public Policy Research, an institute in Wellington.

"You might learn that the money from the award itself makes a big difference, but that the ranking of specific proposals was close to random," says Jaffe, who moved to New Zealand last spring after nearly 2 decades as a professor and dean at Brandeis University in Waltham, Massachusetts. "That would mean it's important for NIH to continue funding

Published by AAAS

research, for example, but that maybe the resources used in the selection process aren't being spent effectively."

Lauer emphasizes that his studies do not mean NIH is funding bad research. Nor is he proposing radical changes in the current system, as some have (see sidebar, p. 598). But he hopes the results prod NIH to question some time-honored assumptions about peer review and focus more on ensuring that its awards are yielding the biggest payoff. "The analogy is to a doctor with a bunch of sick patients," he says. "How do I maximize their health?"

Faith in the system

Ask a scientist about peer review, and many will immediately cite Winston Churchill's famous description of democracy—"the worst form of government except all those other forms that have been tried." The comparison acknowledges the system's many flaws, including its innate conservatism and its inability to make fine distinctions, while providing a defense against attacks from both colleagues and those outside the scientific community. "CSR lives and dies by the belief that our reviews are fair, and that our only bias is around good science," Nakamura says. "And any evidence that suggests otherwise is very troubling."

That's not to say that Nakamura and his colleagues think the current system can't be improved. Last year, Collins asked a group of senior administrators to examine ways of "optimizing peer review" at NIH. In particular, the task force is looking at whether NIH needs to do more to identify and support proposals from emerging fields and, at the same time, learn how to pull the plug on once-hot areas where scientific interest has cooled. "Does the current structure perpetuate fields beyond their prime?" asks Lawrence Tabak, NIH principal deputy director and chair of the task force.

The 170 or so study sections that CSR manages are the essential element of the NIH peer-review system for external grants. Each consists of 12 to 22 outside scientists who meet three times a year to review an average of 70 applications. (Individual institutes also convene review panels of their own.) Panel members give each proposal a numerical score, and the proposal receives an impact score that is the average of individual votes. For many applications, that score is converted into a percentile ranking.

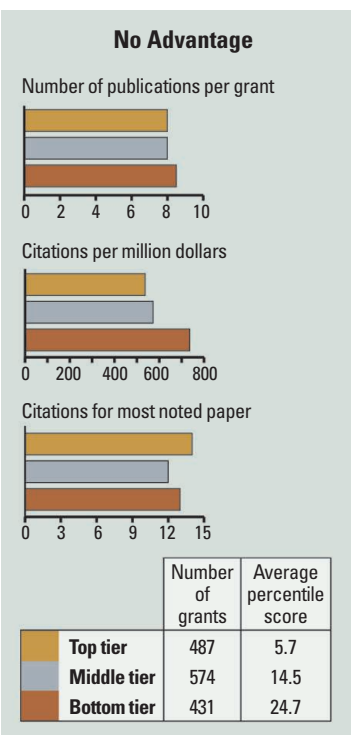
It's a massive system that requires heavy buy-in from the research community. Last year, for example, more than 24,000 scientists reviewed roughly 75,000 applications at some 2500 panel meetings. CSR's budget to manage the entire operation was $110 million.

NIH officials say that peer review is just one building block in constructing a well-balanced portfolio of grants. But they acknowledge that NIH program managers are much more likely than their counterparts at other federal agencies to worry about the consequences of funding a grant "out of order." The assumption is that study sections know best and that a panel's judgment should be overruled only for compelling reasons.

Still, Nakamura is always looking for fresh ways to assess the performance of study sections. At the December meeting of the CSR advisory council, for example, he and Tabak described one recent attempt that examined citation rates of publications generated from research funded by each panel. Those panels with rates higher than the norm—represented by the impact factor of the leading journal in that field—were labeled "hot," while panels with low scores were labeled "cold."

"If it's true that hotter science is that which beats the journals' impact factors, then you could distribute more money to the hot committees than the cold committees," Nakamura explains. "But that's only if you believe that. Major corporations have tried to predict what type of science will yield strong results—and we're all still waiting for IBM to create a machine that can do research with the highest payoff," he adds with tongue in cheek.

"I still believe that scientists ultimately beat metrics or machines. But there are serious challenges to that position. And the question is how to do the research that will show one approach is better than another."



	Number of grants	Average percentile score
Top tier	487	5.7
Middle tier	574	14.5
Bottom tier	431	24.7

Equal impacts. The publication record from proposals with the best scores was no better than for those scoring in the middle and lowest tiers among heart research funded by NIH.

Bolder fixes wanted

Jaffe says he tried for more than a decade to interest top officials at both NIH and the National Science Foundation (NSF) in conducting the same type of quantitative analysis of peer review across the entire agency that Lauer has done within his program. But he was thwarted. "It's amazing to me how scientists who believe in the scientific method don't believe it should be applied to study what they do," Jaffe says. "It's just so intuitively obvious to them that [the current system of peer review] is the best way to do things."

Lauer set out to study peer review within his institute on his own, without funding from CSR. His first inkling that NIH peer review might be falling short came from reading a January 2012 study in *BMJ*. It reported that fewer than half of the researchers funded by NHLBI to test ways of preventing and treating heart disease had published their results within 30 months of the end of the trial, and that one-third of the trials never saw the light of day. A disbelieving Lauer expanded the *BMJ* sample and did his own analysis to be sure. "But they were right," he says. A long lag time might be expected for negative results, but Lauer found the record was no better for studies that came up positive.

He wondered whether the sluggish publication record was a sign of a deeper problem: Maybe the institute was failing to fund studies that journals viewed as urgent and important. And that led him to question whether the peer-review system was doing its job.

So Lauer examined whether funded proposals that were ranked higher during peer review ended up having their data published faster. They didn't. "[T]here was no significant association between the peer-review priority scores received before the trial was funded and the time to publication," he and his NHLBI colleagues wrote last fall in *The New England Journal of Medicine*.

Lauer believes that his findings reflect a bias among reviewers at NHLBI against so-called pragmatic studies, which aim to inform patient care directly by studying a procedure or drug in a typical clinical setting such as an urban hospital. "The message from a pragmatic trial is that you either will or will not do something," he says. "But fewer than 20% of our trials focus on that."

The majority of NIH clinical trials are

SOURCE: ADAPTED FROM DANTHI *ET AL.*, *CIRCULATION RESEARCH* (ADVANCED ONLINE EDITION) 2014

Making Every Scientist a Research Funder

When it comes to using peer review to distribute research dollars, Johan Bollen favors radical simplicity.

Over the years, many scientists have suggested that the current system could be improved by changing the composition of the review panels, tweaking the interactions among reviewers, or revising how the proposals are scored. But Bollen, a computer scientist at Indiana University, Bloomington, would simply award all eligible researchers a block grant—and then require them to give some of it away to colleagues they judge most deserving.

That radical step, described in a paper Bollen and four Indiana colleagues recently posted on *EMBO Reports*, retains peer review's core concept of tapping into the views of the most knowledgeable researchers. But it would eliminate the huge investment in time and money required to submit proposals and assemble panels to judge them.

Bollen's process would be almost instantaneous: In a version of expert-directed crowdsourcing, scientists would fill out a form once a year listing their favored researchers, and a predetermined portion of their annual grant money—a total of, say, 50%—would then be transferred to their choices.

"So many scientists spend so much time on peer review, and there's a high level of frustration," Bollen explains. "We already know who the best people are. And if you're doing good work, then you deserve to receive support."

Others are skeptical. "I've known Johan for a long time and have the highest regard for his ability as an out-of-the-box thinker," says Stephen Griffin, a retired National Science Foundation (NSF) program manager who's now a visiting professor of information sciences at the University of Pittsburgh in Pennsylvania. "But there are a number of issues he doesn't address."

Those sticking points include the likely mismatch between what researchers need and what their colleagues give them; the absence of any replacement for the overhead payments in today's grants, which support infrastructure at host institutions; and the dearth of public accountability for the billions of dollars that would flow from public coffers to individuals. "Scientists aren't really equipped to be a funding agency," Griffin notes.

Bollen acknowledges that the process would need safeguards to ensure that scientists don't reward their friends or punish their enemies. But his analysis suggests that the U.S. research landscape would not look all that different if his radical proposal were adopted.

Drawing upon citation data in 37 million papers over 20 years, the Indiana researchers conducted a simulation premised on the idea that scientists would reallocate their federal dollars according to how often they cited their peers. The simulation, he says, yielded a funding pattern "similar in shape to the actual distribution" at NSF and the National Institutes of Health for the past decade—at a fraction of the overhead required by the current system.

–JDM

instead aimed at testing the underlying biological mechanism of a disease or a treatment. That approach requires researchers to examine a strictly defined set of participants in a specialized clinical environment. Study sections view such an approach as more compelling, Lauer says. "When I go on the road to preach pragmatic trials," he says, "I can predict that someone will stand up and say, 'That's all well and good. But if we submit a proposal for a pragmatic trial, it will get killed in review.' "

Lauer is already at work on changing that perception. The heart division has invited researchers to submit proposals for pragmatic trials that address what he calls "important clinical questions." He's also altered the review process in two significant ways: All the proposals will go to a single panel that will only judge pragmatic trials, and its members will consist of "people who know a lot about pragmatic trials and think they are valuable."

Lauer says his second study, which looked at NIH's bread-and-butter R01 grants, highlights another problem. It divided R01 grants awarded by NHLBI into three pools based on their percentile ranking—better than 10th percentile, 10th to 20th, and 20th to 42nd. His sample went back to 2001, when NIH had the resources to fund some of the proposals in that third tier. These days, proposals in that third category have virtually no chance of being funded, so the fact that the research drew as many citations as top-ranked projects suggests that peer reviewers are ruling out a large share of potentially significant research.

The problems of peer review, Lauer says, are those that afflict any system that relies on the judgments of experts. One eye-opener for Lauer was a 2006 book by Philip Tetlock, a psychologist at the University of Pennsylvania, titled *Expert Political Judgment: How Good Is It? How Can We Know?* The book describes how experts do little better than chance in predicting political events and also vastly overrate their prognosticating abilities. Its lessons apply to peer review as well, Tetlock says "There is high-impact research that has been rejected, and low-impact research that has been funded."

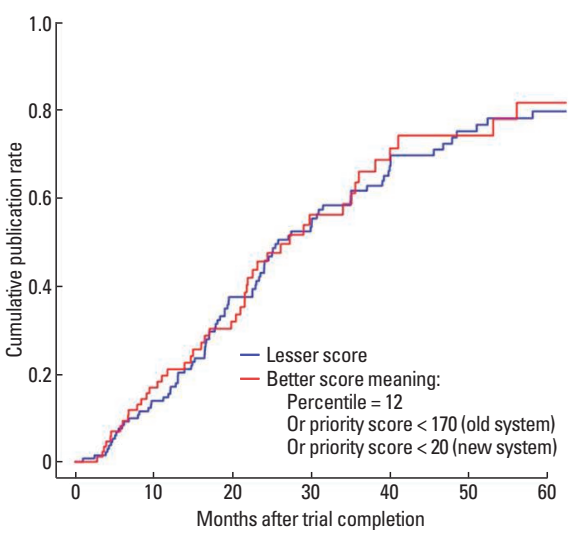
Tetlock is not surprised that NIH hasn't rushed to embrace the kind of self-examination Lauer has done. "Most institutions are not enthusiastic about an objective appraisal of their track records," he says. They are also worried that the results could be used against them. "If the hit rate is not too high and the false positives are high," he says, "people may jump to the conclusion that you guys are a bunch of idiots. In fact, the agency could be doing as good a job as possible, given the unpredictability of the task."

Lauer and Jaffe say NIH should be bolder in designing experiments to improve peer review without abandoning it. In particular, they would like NIH to rigorously test critical components of the system. Possibilities include using a second set of reviewers as a control, or asking reviewers to score proposals on several specific criteria and then tallying up those subscores rather than asking for one overall rating, as is done now.

No system has a 100% hit rate on high-impact programs and never funds a low-impact program, Tetlock notes. "That would be God," he says. But he believes there is plenty of room for improvement. "By using the best science we have on how to elicit and aggregate judgments, maybe we can get our hit rate up and our false positives down," Tetlock says. "And that would be a better world."

–JEFFREY MERVIS

Dead Heat



Slow off the mark. The time to publication for NHLBI clinical trials isn't linked to what review panels thought about the research.

CREDIT: ADAPTED BY M. LAUER FROM D. GORDON *ET AL.*, *NEJM* **369** (14 NOVEMBER 2013)

Published by AAAS