Encouraging high-risk high-reward research at NIH

Paula Stephan, Andrew Young School of Policy Studies, Georgia State University & NBER

Chiara Franzoni,[1] School of Management, Polytechnic University of Milan

May 2023

Note: *This piece was prepared for the Building a Better NIH project, a partnership between the Brookings Institution, the Institute for Progress, and the Good Science Project. For more information about the project, please visit the Building a Better NIH project page.*

A number of scientists have expressed concern that the National Institutes of Health (NIH) is risk averse and becoming increasingly so (Harris, 2013; Lee, 2007). Some think the heavy emphasis placed on demonstrating feasibility of the proposed research by providing preliminary findings is responsible for this risk aversion (Petsko, 2011). More generally, there is the perception that the probability of success plays an important role in the evaluation process (see e.g., Lipinski et al., 2009).

Two studies have found evidence that this may be the case. Azoulay and colleagues (2011) created a matched sample between NIH grant recipients and Howard Hughes Medical Institute (HHMI) recipients and find the HHMI investigators use more novel keywords and produce more hits and more flops, compared with the NIH investigators. It is not clear whether the results depend on the criteria for selection or on other factors, such as the longer duration of grants and the practice of HHMI to not demand early results, but the results are consistent with NIH being more risk averse than HHMI.

A recent article published in PNAS by Packalen and Bhattacharya (2020) examines a time-series of biomedical articles, distinguishing those supported by NIH from those not supported. The authors identify for each paper the *vintage* of the idea, i.e., the year in which the idea that constitutes the main input in the paper first appeared in the literature. They find that, since the 2000s, the share of papers supported by NIH stands in an inverted-U shape with respect to idea vintage (Figure 1, right). This was not the case in the 1990s (Figure 1, left). Overall, the results suggest that NIH has become less supportive of “edgy” research, consistent with risk aversion in selection.

[1] Chiara Franzoni received grant support from the Alfred P. Sloan Foundation and Novo Nordisk Foundation for this work. Other than the aforementioned, the author did not receive financial support from any firm or person for this article or from any firm or person with a financial or political interest in this article. The author is not currently an officer, director, or board member of any organization with a financial or political interest in this article.

Paula Stephan did not receive financial support from any firm or person for this article or from any firm or person with a financial or political interest in this article. The author is not currently an officer, director, or board member of any organization with a financial or political interest in this article.

Although the evidence is based on proxies for risk, not direct measures of risk, it suggests that risky research is disfavored in the competition for NIH funding. It is, however, challenging to test the extent to which NIH is risk averse given that researchers studying the issue only have access to funded proposals; unfunded proposals are not available (NIH, 2022). There is the added challenge of measuring risk.[2]

Proposals to increase the amount of risky research generally fall into two categories (Lipinski et al., 2009): (1) create programs especially designed and targeted to funding risky research; (2) undertake reform of the peer review system for evaluating grants, to make it less risk averse. In this essay we examine these two approaches for fostering funding of more risky research at the NIH. Our focus is on investigator-initiated research undertaken by university faculty. The newly formed Advanced Research Projects Agency for Health (ARPA-H) initiative, recently established at NIH and based on the Defense Advanced Research Projects Agency (DARPA) model, is not discussed here because the model is not directed at funding investigator-initiated research.

To date, the NIH has primarily adopted the first approach for supporting risky research by establishing four awards, depending in part upon the career stage of the applicant, with the goal of fostering "high-risk, high-reward" (HRHR) research. Funds for the programs come out of the Common Fund and the programs are located in the Director's Office. Preliminary data can be included as part of the proposal but are not "expected or required" (NIH, 2023). The Director's Pioneer Award is one of the four; the others are the Director's New Innovator Award, the Director's Transformative Research Award, and the Director's Early Independence Award. The programs are small in scope. For example, approximately seven Pioneer awards are made per year, each for a duration of five years, with annual direct costs of $700,000 per award. This represents an annual commitment on the part of NIH of around $37.24 million with indirect costs included, or 0.15% of the extramural NIH funding budget for an average portfolio of 49 Pioneer Awards being supported a year (NIH Director's Pioneer Award, n.d.; Kaiser, 2017; NIH Extramural & Intramural Funding, n.d.).[3] Success rates are approximately 5% and there is concern that the low success rate discourages faculty from applying for the award (NIH Director's Pioneer Award, n.d.; NIH Director's New Innovator Award, n.d.).[4]

The NIH also has the R21 "Exploratory/Developmental Research Grant Award", which provides "support for the early stages of project development." Funds are limited to $275,000 for a two-year period (NIH Director's New Innovator Award, n.d.). Early-career stage faculty can use R21 funds to bolster their first R01 submission (Types of Grant Programs, n.d.; NIA New and Early Stage Investigators, n.d.). NIH receives approximately 50% fewer R21 applications a year than R01 applications and the success rate is slightly lower for R21s than for R01s (Lauer, 2016).

---

[2] One measure that has been used as a proxy for risk is that of "novelty," which measures the degree to which a research paper draws on combinations of references that have, to date, never been combined before (Wang, Veugelers, and Stephan, 2017).

[3] "NIH Director's Pioneer Award" (n.d.). Calculation assumes a 52% indirect rate on average (Kaiser, 2017) and uses an extramural budget of 34,407,000,000 ("NIH Extramural & Intramural Funding," n.d.).

[4] "NIH Director's Pioneer Award" (n.d.). Success rate from 2017-2021 was 4.9%. Considerably more New Innovator awards are made; for example, in 2022, 35 were expected to be awarded ("NIH Director's New Innovator Award," n.d.).

We have no clear evidence to judge if the stated goal of the NIH "high-risk, high reward" programs is successful in promoting more risky research.[5] Our priors are that it likely is not, given that the HRHR programs in the Director's Office constitute but a fraction of all NIH funding. And while there are many more R21 awards, the size and length of the award is unlikely to encourage risky research. Moreover, R21s are used as a preparatory grant to subsequent R01 submission, suggesting that researchers, knowing they will be submitting a R01 application, have a strong incentive to get research "out-the-door" and published.

More generally, there are good reasons to call into question the extent to which funders, despite articulating the goal of supporting high risk research, can achieve a HRHR goal without also undertaking a systemic reform of peer-review evaluations of proposals. In a recent study, Veugelers et al. (2021) examine the funding choices of the European Research Council (ERC), the granting agency established by the European Commission to support exclusively HRHR research. They find that, despite the ERC commitment to fund HRHR science, the applicants with a history of doing novel research were significantly less likely to be funded than were those without such a history.[6]

Although procedures for scoring the R01, the dominant NIH funding vehicle for university faculty, have undergone some revision over time, minimal effort has been made to incorporate risk into the panel's scoring and selection procedures. These procedures currently involve panelists reading and rating a subset of proposals independently from the other reviewers before the panel meets, and then assigning a preliminary impact score that reflects the "overall impact that the project is likely to have on the research field(s) involved".[7] The overall impact scores provided by the reviewers are submitted to the panel prior to the meeting. Selection of proposals to discuss is done by rank-ordering the proposals on the basis of the average overall impact scores given by panel members who reviewed them. The worst-performing are immediately eliminated. The panel then meets to discuss the remaining proposals. After discussion, each voting member provides a final impact score. These are added together and the mean score is computed. Proposals are then arrayed in terms of mean impact score, and funding is recommended starting with the application with the best score until the budget is

[5] Evaluations to date typically focus on application and review process; not on outcomes. See https://commonfund.nih.gov/evaluationlibrary.

[6] The result is robust for those applying to the "starter" program but not to the "advanced" program and is consistent with an aversion to funding early-career researchers with a history of engaging in risky research. Not discussed here is the question of whether applicants actually submit risky projects or, fearing a risk bias on the part of reviewers, play it safe by submitting less risky projects. For a discussion, see https://blog.ninds.nih.gov/2014/03/27/back-to-basics/. Accessed March 30, 2021.

[7] Evaluation is required along five criteria (Significance, Investigator(s), Innovation, Approach, Environment). See webpage: https://grants.nih.gov/grants/guide/notice-files/NOT-HS-10-002.html. Accessed March 22, 2023. The five criteria are scored on a scale [1-9], with 1 being the highest score. Reviewers are also asked to provide comments for each score. See "Scoring guidance" webpage: https://grants.nih.gov/grants/peer/guidelines_general/scoring_system_and_procedure.pdf. Accessed March 22, 2023. Note, that prior to 2006 the strength of preliminary results was also scored. Although the criterion was eliminated, there is widespread concern that the overall score reflects the strength of preliminary results.

depleted although some consideration is given to funding proposals that fall below the "pay line."

There are several aspects of the evaluation process that can contribute to placing risky research at a disadvantage. First, panelists may implicitly place a heavy weight on criteria, such as the "Approach", which emphasizes feasibility or having convincing preliminary results. An analysis of criteria scores conducted at NIH in 2010 supports this view (Rockey, 2011). "No crystal, no grant" was the saying at the time the National Institute of General Medical Science (NIGMS) evaluated protein structure proposals. To the extent this is the case, projects of low-to-medium value with an extremely high probability of success are likely to be selected over projects of high potential value with low-to-medium probability of success. In an experiment conducted at the Harvard Medical School, Boudreau et al. (2016) find that more novel research proposals, as measured by the percent of keywords not previously used, receive more negative evaluations during peer-review. Second, exchange of information among panel members may lead to what is sometimes referred to as "groupthink", which, in the peer-review context, may easily translate to a *negativity bias*. A recent study by Lane et al. (2021), for example, found that panel members tend to lower their score upon learning that another panel member has rated the proposal lower than they have.[8] No evidence was found that panel members raised their score when presented with a higher score by another panel member.

Given that novel research may be more exposed to criticism, selection mechanisms that either do not require the sharing of information or the need to achieve consensus may be more supportive of risky research. The idea of providing a golden ticket, whereby each panel member can select one proposal to fund, regardless of the votes of other panel members, has been proposed as a possible solution to the problem. To date, the Villum Foundation and the Volkswagen Foundation are the only funding organization that have used golden tickets. A randomized control trial has yet to be done. The programs are small. In the case of Villum, 15% of the budget is allocated by golden tickets. In 2022, 51 researchers received 100 million kroner ($14.3 million in USD) through the scheme (Singh Chawla, 2023).

Several other parameters were varied at the same time the golden tickets were using, making interpretation of the results difficult.[9] It is worth noting, however, that approximately one-half of the Villum golden ticket-funded-proposals would not have been funded based on scores, and none of the golden ticket Volkswagen awardees would have been funded based on scores (Sinkjaer, 2018).

Another possible solution is to use a full or partial randomization (or *lottery*) to choose between eligible proposals, rather than rely on the panel decisions. Testing, once again, has been rather limited. For example, the Explorer's Grant of the New Zealand's Health Research Council in 2013 adopted a random number generator to select four innovative and out-of-the-box research projects per year in all areas of sciences (Bendiscioli, 2019; Sinkjaer, 2018). The lottery approach has also been applied as a complement to peer review, for example to select only

[8] Unlike at NIH, in the experiment, lower scores represented that the panel member had a lower opinion of the proposal than did higher scores.
[9] In the case of Villum, for example, the names of the applicants were withheld from reviewers.

among the proposals that are in the middle of the distribution of reviewers' average scores, while reviewers scores are used for selecting at the top and bottom of the distribution. [10] The *partial lottery* approach is currently used in a program for postdoctoral scholarships conducted by the Swiss National Science Foundation (Heyard et al., 2021) and it has been pilot-tested in the past by the Volkswagen Foundation (Bendiscioli, 2019). Although to date virtually no experimental work has analyzed the extent to which the partial lottery fosters the selection of risky research, the argument can be made that, at least for a subset, in theory it should. The logic is that proposals in the middle of the distribution have average scores either because panelists have scored them as average or because some panelists have given them high scores while others have given them low scores. Panel discussion for proposals in the latter situation is decisive and can determine whether or not they are funded. Consensus is arguably key in such situations because if some reviewers are strongly against the proposal, the panel likely will opt for other proposals. A lottery for proposals in the middle of the distribution with a wide variance in scores thus can de-emphasize consensus and plausibly increase the odds that risky proposals are funded.

Golden tickets and lotteries may be possible solutions to overcoming risk biases that may intervene during the final deliberation process by panel members. But risky research can be dismissed even before reaching the stage of panel deliberations. To the extent that reviewers place heavy emphasis on partial aspects of a proposal, such as the requirement of preliminary findings, this may be the case.[11] We recently proposed a way to reform the protocol for peer review with the goal of making it better at assessing HRHR research. Specifically, the proposal is that panel members consider in a systematic and more comprehensive way the risk involved in a proposal as depending on the probability that the research will be successful as well as the potential value of the findings, if successful (Franzoni and Stephan, 2023). The core idea is that HRHR proposals generally can be thought of as i) entailing a very uncertain distribution of outcomes; and ii) being more explorative in nature, thus more prone to fail, but also to lead to important outcomes, often beyond the main aim of a research proposal. In light of this, the existing protocols of grant peer review have two major pitfalls. First, by asking for a single score, they do not elicit the probabilistic nature of outcomes and scoring may be dominated by concern that the research approach has not been proven and involves high chances of failure. This often leads to entirely ignoring the small chances of success. Second, they do not ask (nor encourage) reviewers to explore the entire distribution of the possible outcomes, especially the positive spectrum of the outcome distribution. Specifically, reviewers are not instructed to weigh the possibility that the research leads to important, albeit unlikely, findings, including the possibility of important secondary outcomes. (Franzoni and Stephan, 2023).

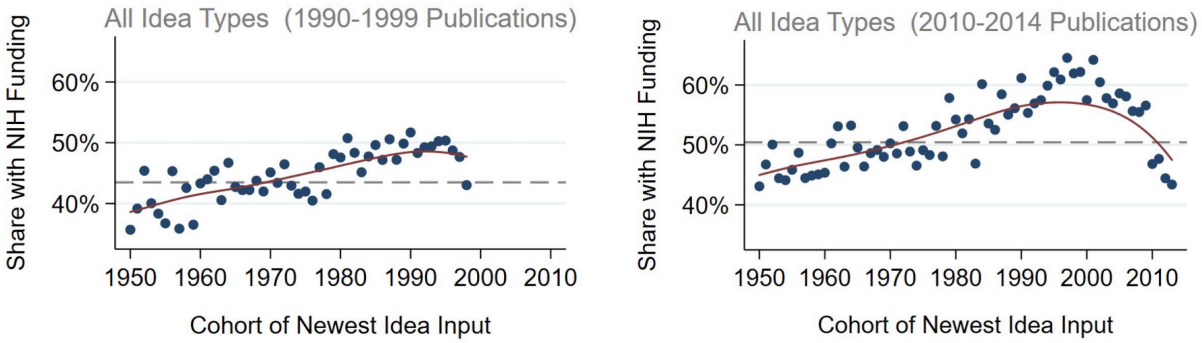
The approach we propose (Franzoni and Stephan, 2023) draws on subjective expected utility. In its most basic form, it asks evaluators to reflect and assess first the value that a proposed primary outcome may entail for science and/or society, then assess the probability of the primary outcome occurring. Evaluators are asked to do the same for secondary findings if these are envisaged. Once the evaluators provide these inputs, there is no need to ask for an

[10] Proposals at the top of the distribution are funded; proposals at the bottom are rejected.
[11] Note that individual scores and initial scores of funded proposals are not available at NIH; only the final score is provided (Li, 2017).

additional overall score. The values provided are sufficient for computing a numeric score that represents the sum of values times probability across the possible combinations of primary and secondary outcomes.[12] The formula for automatically computing the overall score ensures that all components are considered in the overall evaluation. For example, it addresses the concern that the heavy focus on preliminary findings can lead reviewers to discount research of high potential value, if successful, but with a low probability of success, in favor of proposals of low to medium potential value but a high probability of success. The approach we propose is applicable to programs such as the R01 proposals, but it is also applicable to the funding programs especially designed for HRHR.

The proposed approach has yet to be used in the peer review evaluation process. If it were to be considered, it would be important to design experiments to see the extent to which the approach yields different outcomes then the current NIH procedure does and the extent to which it funds riskier research. Other approaches, such as a lottery or a golden ticket are also available. They, too, could be evaluated by designing experiments. Our goal here is not to "solve" the apparent risk bias of NIH but to raise it as a concern and propose possible ways to address the bias.

Figure 1. Share of NIH-funded publications over total by vintage of idea



Share of total publications funded by NIH by vintage of the newest idea in publication.
Left: 1990-1999 publications. Right: 2010-2014 publications.
Markers: mean NIH funding rate. Dashed line: average funding rate across all vintages. Solid line: non-parametric regression estimate. Source: Packalen & Battacharya, PNAS 2020, p.12013.

[12] Specifically, the Subjective Expected Utility (SEU) of the project *i* (Y_i) is the sum of values ($u_{(.)}$) times probability ($P_{(.)}$) across the possible combinations of primary and secondary outcomes, i.e. $Yi = u1i \cdot P1i + u2i \cdot P2i + (u1i + u2i) \cdot P2i \cdot P1i$ (Franzoni and Stephan 2023).

References

Azoulay, Pierre, Graff Zivin, J.S., & Manso, G. (2011). Incentives and Creativity: Evidence from the Howard Hughes Medical Investigator Program. The Rand Journal of Economics 42(3):527–54. https://onlinelibrary.wiley.com/doi/10.1111/j.1756-2171.2011.00140.x

Bendiscioli, S. (2019). The Troubles with Peer Review for Allocating Research Funding. EMBO Reports 20(12). https://www.embopress.org/doi/full/10.15252/embr.201949472

Boudreau, K. J., Guinan, E.C., Lakhani, K.R., & Riedl, C. (2016). Looking beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. Management Science, 62(10). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2478627

Franzoni, C. & Stephan, P. (2023). Uncertainty and Risk-Taking in Science: Meaning, Measurement and Management in Peer Review of Research Proposals. Research Policy; 52(3). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3804560

Guthrie, S., Ghiga, I., & Wooding, S. (2018). What Do We Know about Grant Peer Review in the Health Sciences? An Updated Review of the Literature and Six Case Studies. RAND Corporation. https://doi.org/10.7249/RR1822

Harris, R. (2013). Scientists Win Nobel For Work On How Cells Communicate. NPR. Accessed November 18, 2022. https://www.npr.org/templates/story/story.php?storyId=230192033.

Heyard, R., Ott, M., Salanti, G., & Egger, M. (2022). Rethinking the Funding Line at the Swiss National Science Foundation: Bayesian Ranking and Lottery. Statistics and Public Policy, 9(1). https://www.tandfonline.com/doi/full/10.1080/2330443X.2022.2086190

Kaiser, J. (2017). NIH plan to reduce overhead payments draws fire. Science. Accessed March 23, 2023. https://www.science.org/content/article/nih-plan-reduce-overhead-payments-draws-fire

Lane, J.N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E.C., & Lakhani, K.R. (2021). Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation. Management Science 68(6): 4478-4495. https://pubsonline.informs.org/doi/10.1287/mnsc.2021.4107

Lauer, M. (2016). R01 and R21 Applications & Awards: Trends and Relationships Across NIH. Office of Extramural Research, National Institutes of Health. Accessed October 4, 2022. https://nexus.od.nih.gov/all/2016/11/04/nih-r01-r21/

Lee, C. (2007, May 28). Slump in NIH Funding Is Taking Toll on Research. Washington Post.

Li, D. (2017). Expertise versus Bias in Evaluation: Evidence from the NIH. American Economic Journal: Applied Economics 9(2):60–92. https://www.aeaweb.org/articles?id=10.1257/app.20150421

Lipinski, D., Ehlers, V.J., Lane, N.F., Collins, J.P., McCullough, R.D., Rubin, G.M., & Orr, F.M.. (2009). Investing in High-Risk, High-Reward Research. 111th Cong. Washington: House of Representatives.

Granger, M.M. & Henrion, M. (1990). Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge, MA: Cambridge University Press.

NIA New and Early Stage Investigators. (n.d.). National Institute on Aging, National Institutes of Health. Accessed March 23, 2023. https://www.nia.nih.gov/research/grants-funding/nia-new-and-early-stage-investigators

NIH Director's New Innovator Award: Frequently Asked Questions. (n.d.). National Institutes of Health. Accessed March 23, 2023. https://commonfund.nih.gov/newinnovator/faq

NIH Director's Pioneer Award: Frequently Asked Questions. (n.d.). National Institutes of Health. Accessed October 4, 2022. https://commonfund.nih.gov/pioneer/faq

NIH Extramural & Intramural Funding: FY 2021 Operating Plan. (n.d.). NIH Data Book, National Institutes of Health. Accessed March 23, 2023. https://report.nih.gov/nihdatabook/report/283

NIH Grants Policy Statement: 2.3.11 Availability and Confidentiality of Information. (2022). "." National Institutes of Health. Accessed October 4, 2022. https://grants.nih.gov/grants/policy/nihgps/html5/section_2/2.3.11_availability_and_confidentiality_of_information.htm

High-Risk, High-Reward Research. (2023). National Institutes of Health. Accessed October 4, 2022. https://commonfund.nih.gov/highrisk

Packalen, M. & Bhattacharya, J. (2020). NIH Funding and the Pursuit of Edge Science. PNAS 117(22):12011–16. https://www.pnas.org/doi/10.1073/pnas.1910160117

Petsko, G. (2011). Risky Business. Genome Biology 12(119). https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-6-119

Rockey, S. (2011). Correlation Between Overall Impact Scores and Criterion Scores. Office of Extramural Research, National Institutes of Health. Accessed Nov 18, 2022. https://nexus.od.nih.gov/all/2011/03/08/overall-impact-and-criterion-scores/

Singh Chawla, D. (2023). 'Golden tickets' on the cards for NSF grant reviewers. Nature. https://doi.org/10.1038/d41586-023-00579-z

Sinkjaer, T. (2018). Fund Ideas, Not Pedigree, to Find Fresh Insight. Nature 555:143.
Types of Grant Programs. (n.d.). Grants & Funding, National Institutes of Health. Accessed October 4, 2022. https://grants.nih.gov/grants/funding/funding_program.htm#RSeries

Veugelers, R., Stephan, P., & Wang, J. (2021). Excess Risk-Aversion at ERC. Mimeo.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators. Research Policy 46(8):1416–36. https://www.sciencedirect.com/science/article/abs/pii/S0048733317301038?via%3Dihub