Predicting Productivity Returns on Investment
Thirty Years of Peer Review, Grant Funding, and Publication of Highly Cited Papers at the National Heart, Lung, and Blood Institute

Michael S. Lauer, Narasimhan S. Danthi, Jonathan Kaltman, Colin Wu

There are conflicting data about the ability of peer review percentile rankings to predict grant productivity, as measured through publications and citations. To understand the nature of these apparent conflicting findings, we analyzed bibliometric outcomes of 6873 de novo cardiovascular R01 grants funded by the National Heart, Lung, and Blood Institute (NHLBI) between 1980 and 2011. Our outcomes focus on top-10% articles, meaning articles that were cited more often than 90% of other articles on the same topic, of the same type (eg, article, editorial), and published in the same year. The 6873 grants yielded 62468 articles, of which 13507 (or 22%) were top-10% articles. There was a modest association between better grant percentile ranking and number of top-10% articles. However, discrimination was poor (area under receiver operating characteristic curve [ROC], 0.52; 95% confidence interval, 0.51–0.53). Furthermore, better percentile ranking was also associated with higher annual and total inflation-adjusted grant budgets. There was no association between grant percentile ranking and grant outcome as assessed by number of top-10% articles per $million spent. Hence, the seemingly conflicting findings on peer review percentile ranking of grants and subsequent productivity largely reflect differing questions and outcomes. Taken together, these findings raise questions about how best National Institutes of Health (NIH) should use peer review assessments to make complex funding decisions.

A just-published analysis by Li and Agha1 of ≈30 years of NIH R01 grants showed associations between better percentile rankings and bibliometric outcomes. These associations persisted even after accounting for several potential confounding variables, including prior investigator track record and institutional funding. These associations also seem to be at odds with prior analyses from the NHLBI,2–4 the National Institute of General Medical Sciences,5 the National Institute of Mental Health,6 and the National Science Foundation.7 How can we reconcile these apparent differences? Are these differences contradictory or do they reflect questions that differ in a subtle, although important manner?

To understand the different findings, it is important to consider the differences between Li and Agha1 and the priory reports. The most obvious, perhaps, is that Li and Agha1 included a much larger number of grants that were funded for many decades. But there are 2 other key differences: first, Li and Agha focused on raw publication and citation counts, as opposed to field normalized counts,8 and second Li and Agha focused on bibliometric outcomes alone, whereas some of the previous studies focused on outcomes per $million spent.2–4,6

If you were told that a person weighs 100 pounds, you would know little. If you were then told that a person is a 6-foot tall man, we might worry about cachexia. If you were told that a person is a 10-year-old girl, we would worry about serious obesity. Similarly, if you were told that an article received 100 citations, you would know little. Your interpretation would change depending on whether the article focuses on mathematics, cell biology, basic cardiovascular biology, or clinical cardiovascular medicine.9 It would also change if the article was published 1 year ago or 10 years ago. One recent analysis found that clinical cardiovascular articles are cited 40% more often than basic articles, and that citation rates in cardiovascular sciences have increased dramatically over time.10 Because of these marked variations in citation practice, several authorities9 identify the percentiles approach as the most robust citation metric.11 Here, each article is judged against other articles published in the same year and dealing with the same topic; hence a biochemistry research article published in 2005 is compared against other biochemistry research articles published in 2005 and not against a clinical trial article published in 2002.

Another question is whether one measures outcome alone or outcomes in light of money spent. Every grant or contract that NHLBI dispenses incurs opportunity costs; if NHLBI chooses to fund a large, expensive trial, that means it will not be able to fund a certain number of smaller (in terms of budget) R01 grants. If we focus on bibliometric outcomes, certainly not the only outcomes worth considering, we would not ask how many highly cited (for field and year) articles were produced, but how many were produced for every $million spent?
spent.\textsuperscript{4} In other words, the outcome metric for the previous studies was not only return but also return on investment.

To gain greater insight into the seemingly different outcomes of Li and Agha and prior reports, we now turn to examine bibliometric outcomes of 30 years of cardiovascular grants funded by the NHLBI.

Between 1980 and 2011, NHLBI funded 8125 de novo (ie, not renewals) cardiovascular R01 grants. Of these 6873 were investigator-initiated and received a percentile ranking, whereas the remainder did not receive a percentile ranking mainly because they were reviewed by ad hoc panels assembled in response to specific programs. Of these 6873 percentiled grants, 1867 (27\%) were successfully renewed at least once. Through 2013, these percentiled grants generated 62468 articles; of these, 13507 (or 22\%) were top-10\% articles, meaning that they were cited more often than 90\% of all other articles published in the same year and focused on the same topic. The expected value would be 10\%, meaning that as a whole the portfolio performed at least twice as well as would be expected by chance.

The distribution of top-10\% articles was highly skewed, consistent with prior work showing the heavy-tailed nature of scientific output.\textsuperscript{12} That is, a small proportion of scientific effort is responsible for a disproportionately large proportion of output; in common parlance, there is a 20 to 80 phenomenon in which 20\% of the input yields 80\% of the output. The median number of top-10\% articles per grant was 1 (interquartile range, 0–2) with a range of 0 to 154. Because of the skewed distribution, we show all analyses after logarithmic transformation.

Figure 1 shows the association of top-10\% articles and percentile ranking. Consistent with Li and Agha,\textsuperscript{1} grants with better percentile scores generated more top-10\% articles. However, the individual points, each referring to 1 grant, shown in the top panel, illustrate the high degree of noise. To assess how well percentile ranking discriminated between grants more or less likely to generate top-10\% articles, we calculated the area under a receiving operator curve (area under the curve under a ROC; Online Figure I) and found modest discrimination that was only slightly better than chance (area under the curve, 0.52; 95\% confidence interval, 0.51–0.53). Grants with better percentile scores had higher annual and total inflation-adjusted budgets (Figure 2, left), and grants with higher annual and total budgets yielded more top-10\% articles (Figure 2, right), although with varying marginal returns. The association between better percentile scores and higher budgets is not only a reflection of actual allocations (Figure 2, left) but also of requested allocations (Online Figure II; based on a separate set of grant applications submitted in 2011–2012). Sometimes, allocated budgets are lower than requested budgets, usually because of postreview negotiations between program staff and applicants.

Budget, of course, is a critical component because an NIH institute should not only be concerned about return (in this case number of top-10\% articles), but about return on investment (number of top-10\% articles per $million spent). Grants with better percentile rankings may generate more top-10\% articles, but also tend to also cost the NIH more money.

Figure 3 shows the association between top-10\% articles per $million spent and percentile ranking. Consistent with prior reports,\textsuperscript{2–4,6} admittedly based on smaller samples studied during shorter periods of time, there was no association between grant percentile ranking and top-10\% articles per $million spent. We also found no ability of percentile ranking to discriminate grants with higher and lower productivity by this metric (area under the curve under the ROC, 0.49; 95\% confidence interval, 0.47–0.50; Online Figure III).

Thus, it seems that some of the apparent contradiction between the recent analysis of Li and Agha\textsuperscript{1} and the prior ones\textsuperscript{2–6} is that each focused on different questions. To maximize return on investment, it may be more appropriate for NIH to focus on bibliometric outcomes per $million spent rather than on bibliometric outcomes alone. We must acknowledge, although, that this metric has its limitations; that is, a metric like number of highly cited articles per $million may not be able to fully discriminate which projects yielded greatest scientific impact. In some cases it may make sense to spend more even for a small number of highly cited articles; for example, NIH
may wisely choose to spend proportionally more money to fund clinical trials that directly affect practice and improve public health. Other projects may generate new methods that are well worth the money because they make whole new areas of scientific inquiry possible. Citation metrics have well-known limitations; for example, articles may be cited because they are controversial rather than meritorious. Nonetheless, recent literature has found a correlation between expert opinion and citation impact. Furthermore, assessments of return on investment should also consider other factors, such as the type of research. We recently reported that at the National Institute of Mental Health, basic science projects seem to yield a greater return on investment than applied projects; we are planning similar analyses at NHLBI.

What can we say from all this? It seems that peer review is able, to a modest extent and with modest degrees of discrimination, to predict which grants will yield more top-10% articles. The modest degree of discrimination reflects previous reports suggesting that most grants funded by a government sponsor are to some extent chosen at random. At the same time, it seems that this association is closely entangled with budgets, budgets that NIH institutes must wrestle with. Because of the diminishing marginal returns seen with more expensive, and better scoring, grants, these data challenge us to question the assumption that it is best to rely primarily on the payline for making funding decisions. Is it smart investment strategy to fund all grants below the cutoff payline at close to requested budgets, but fund a tiny fraction of grants for those scoring above the payline?

In an important sense, we are not so much dealing with a peer review question, but a larger systems question: how best should NIH make funding decisions about those grants that pass the muster of peer review? NIH might ask reviewers to address explicitly their perspectives on the opportunity costs of applications, especially those that are more expensive. Instead of making funding decisions solely based on peer review rankings, decision makers could consider several alternate approaches that leverage, but do not solely rely on, peer review. Some NIH institutes have already taken explicit steps to move away from strict adherence to paylines, choosing instead to fund a proportion of grants from among those that score generally well, that is within a zone of opportunity. Some thought leaders have called for NIH to pay closer attention to the distribution of funding, thereby enabling more scientists to benefit from increasingly constrained funds. Some institutes are choosing to change the

Figure 2. Loess smoother and 95% confidence ranges for the association between budgets, grant percentile ranking, and number of top-10% articles generated per grant. Budget values are shown in constant 2000 dollars (after logarithmic transformation) with inflation-adjustments according to the Biomedical Research and Development Price Index (BRDPI). Top left, Focuses on annual budgets; bottom left, focuses on total budgets (which reflect annual budgets and project duration). The 2 right-sided panels show the association between the number of top-10% articles generated per grant and budgets awarded over a grant’s lifetime (including renewals when applicable). Both x and y axes are logarithmically transformed. The dotted line represents a slope of 1, where a 20% increase in budget would be associated with a 20% increase in the number of top-10% articles. Top right, Focuses on annual budgets; bottom right, focuses on total budgets (which reflect annual budgets and project duration). Bottom right, The actual curve is beneath the slope-of-one line, consistent with diminishing marginal returns.
focus of peer review from a system that primarily focuses on the merit of projects to one that focuses on the expected performance of people and their research teams.\textsuperscript{17} Still another approach depends on the type of research; for example, when NIH considers which clinical trials to fund it might choose to prioritize those that focus on hard clinical end points and to take steps to assure that peer review panels appreciate the importance of such trials.\textsuperscript{18,19} These alternate approaches—the zone of opportunity, additional scrutiny for well-funded investigators, focus on people instead of projects, and NIH stipulation of preferred clinical trials—are only a subset of possibilities, which all beg for their own rigorous analyses to determine whether they enable NIH to make better decisions.\textsuperscript{20} Some critics have argued that it is not enough for NIH to try new and different mechanisms: NIH should turn the scientific method on itself by conducting its own randomized trials,\textsuperscript{21,22} some of which could well involve peer review and how NIH program staff respond to peer reviewers’ scores and comments. In any case, the recently burgeoning literature on grant peer review promises to usher in a period in which NIH, the scientific community, and the public will engage in a rich, data-driven dialogue on how best to leverage the scientific method to improve public health.

**Data and Methods to Generate Plots**

Data on NHLBI cardiovascular R01 grants, including data on percentile rankings and budgets, were obtained from internal tables. All budget figures were inflation-adjusted to 2000 constant dollars using the Biomedical Research and Developments Price Index (or BRDPI at [http://officeofbudget.od.nih.gov/gbiPriceIndexes.html](http://officeofbudget.od.nih.gov/gbiPriceIndexes.html)). We used the SPIRES (Scientific Publication Information Retrieval and Evaluation System) system to match grant numbers to publication identifiers (PubMed Identification [PMIDs]) and supplemented PMID with bibliographic data stored in an EndNote library. We worked with Thomson Reuters to link these publications to their InCites database, which generated for each publication a percentile value describing how often that publication was cited compared with similar publications on the same topic, of the same type (eg, article, letter, editorial), and in the same year. The InCites database empirically classifies articles according to 252 distinct topics; among the 62,468 articles considered here, the most common topics were cardiac and cardiovascular systems, biochemistry and molecular biology, physiology, peripheral vascular disease, and pharmacology and pharmacy. As we have described previously, we divided credit for articles if they cited >1 grant in the portfolio; thus, if a article was classified as a top-10% article and it acknowledged 2 grants, each grant was credited with 0.5 top-10% articles. We generated scatter plots, loess smoother, and 95% confidence intervals with Wickham’s ggplot2 R package.\textsuperscript{23} We calculated areas under receiver operating curves using the pROC R package\textsuperscript{24} and plotted the curves with the plotROC package.\textsuperscript{25}

**Acknowledgments**

We are grateful to Gary Gibbons (Director of National Heart, Lung, and Blood Institute [NHLBI]), Jon Lorsch (Director of National Institute of General Medical Sciences), and Richard Hodes (Director of National Institute on Aging) for their helpful comments on earlier versions of this article. We are grateful to Ebyan Addou and Sean Coady for curating grant and citation data. We also thank Donna DiMichele for helping NHLBI secure InCites data. The views expressed here are those of the authors and do not necessarily reflect the view of the NHLBI, the National Institutes of Health, or the Federal government.

**Disclosures**

We are full-time National Institutes of Health employees and conducted this work as part of our official federal duties.

**References**


15. Graves N, Barnett AG, Clarke P. Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel. BMJ. 2011;343:d4797.


KEY WORDS: bibliometrics ■ National Institutes of Health (US) ■ National Heart, Lung, and Blood Institute (US) ■ peer review ■ ROC curve
Predicting Productivity Returns on Investment: Thirty Years of Peer Review, Grant Funding, and Publication of Highly Cited Papers at the National Heart, Lung, and Blood Institute

Michael S. Lauer, Narasimhan S. Danthi, Jonathan Kaltman and Colin Wu

Circ Res. 2015;117:239-243; originally published online June 18, 2015;
doi: 10.1161/CIRCRESAHA.115.306830

Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2015 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circres.ahajournals.org/content/117/3/239

Data Supplement (unedited) at:
http://circres.ahajournals.org/content/suppl/2015/06/18/CIRCRESAHA.115.306830.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation Research can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation Research is online at:
http://circres.ahajournals.org/subscriptions/
Supplementary Material for:

Predicting Productivity Returns on Investment: Thirty Years of Peer Review, Grant Funding, and Publication of Highly Cited Papers at the National Heart, Lung, and Blood Institute

Michael S Lauer, Narasimhan S. Danthi, Jonathan Kaltman, and Colin Wu

All from the Division of Cardiovascular Sciences, NHLBI
Online Figure I: Receiver operating characteristic curve of the ability of grant percentile ranking to predict a grant’s publication of at least one top-10% paper. A false positive would occur if a grant with a percentile ranking below a certain cut-off failed to produce at least one top-10% paper, whereas a true positive would occur should such a paper be published.

AUC 0.521 (95% CI 0.507–0.534)
Online Figure II: Loess smoother and 95% confidence ranges for the association between requested total budgets and grant percentile ranking among 3101 de novo R01 applications submitted to NHLBI in 2011 and 2012.
Online Figure III: Receiver operating characteristic curve of the ability of grant percentile ranking to predict a grant's publication of at least one top-10% paper per $million spent. A false positive would occur if a grant with a percentile ranking below a certain cut-off failed to produce at least one top-10% paper per $million spent, whereas a true positive would occur should such a paper be published.

AUC 0.489 (95% CI 0.474–0.503)