PLOS ONE

RESEARCH ARTICLE

Examining the Predictive Validity of NIH Peer Review Scores

Mark D. Lindner*, **Richard K. Nakamura**

Center for Scientific Review, National Institutes of Health, 6701 Rockledge Dr., Bethesda, Maryland, United States of America

* Mark.Lindner@nih.gov

Abstract

The predictive validity of peer review at the National Institutes of Health (NIH) has not yet been demonstrated empirically. It might be assumed that the most efficient and expedient test of the predictive validity of NIH peer review would be an examination of the correlation between percentile scores from peer review and bibliometric indices of the publications produced from funded projects. The present study used a large dataset to examine the rationale for such a study, to determine if it would satisfy the requirements for a test of predictive validity. The results show significant restriction of range in the applications selected for funding. Furthermore, those few applications that are funded with slightly worse peer review scores are not selected at random or representative of other applications in the same range. The funding institutes also negotiate with applicants to address issues identified during peer review. Therefore, the peer review scores assigned to the submitted applications, especially for those few funded applications with slightly worse peer review scores, do not reflect the changed and improved projects that are eventually funded. In addition, citation metrics by themselves are not valid or appropriate measures of scientific impact. The use of bibliometric indices on their own to measure scientific impact would likely increase the inefficiencies and problems with replicability already largely attributed to the current over-emphasis on bibliometric indices. Therefore, retrospective analyses of the correlation between percentile scores from peer review and bibliometric indices of the publications resulting from funded grant applications are not valid tests of the predictive validity of peer review at the NIH.

Citation: Lindner MD, Nakamura RK (2015) Examining the Predictive Validity of NIH Peer Review Scores. PLoS ONE 10(6): e0126938. doi:10.1371/journal.pone.0126938

Academic Editor: Neil R. Smalheiser, University of Illinois-Chicago, UNITED STATES

Received: February 23, 2015

Accepted: March 30, 2015

Published: June 3, 2015

Data Availability Statement: All relevant, de-identified data are uploaded to the NIH web site for the Center for Scientific Review at the following URL: http://public.csr.nih.gov/aboutcsr/NewsAndPublications/News/Documents/DEIDENTIFIEDDATASETUSINGZSCORES.XLSX.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The National Institutes of Health (NIH) supports the training, development and research of more than 300,000 full-time scientists working throughout the United States, which stimulates economic activity, produces new businesses and products, and advances basic and clinical biomedical research [1,2]. The NIH has supported the majority of landmark studies that led to Nobel prizes [3], and more than 67% of all citations in scientific articles refer to studies funded primarily by the NIH [4]. In addition, every $1 of NIH funding produces $2.2 – $2.6 or more

in economic activity [1,2,5], more than half of all the studies cited in patents filed in the biomedical field were funded primarily by the NIH [4,6,7], and many of the new drugs and biologics produced are based on research funded by the NIH [8,9], all of which contribute to significant increases in the length and quality of life [10,11].

Since the end of WWII, the NIH has largely relied on a peer review process to allocate extramural funds [12]: scientists submit applications for funding, and other scientists who are active in the same fields—their peers—are recruited by the NIH to review those applications and determine the quality, scientific merit and potential impact of the research described in those applications. Despite clear historical evidence that NIH-funded research has produced a wide range of significant and beneficial effects, the scientific community is expressing increasing criticism of the peer review process that the NIH relies on to allocate funds [13–15], and in fact, the predictive validity of peer review has not yet been empirically demonstrated [16].

Perhaps the most efficient and expedient test of the predictive validity of NIH peer review would be a retrospective analysis of the correlation between percentile scores from peer review as the predictor and bibliometric indices as the criteria. The percentile score is used by each institute as the primary influence when deciding which applications to fund, and scientists are evaluated by their academic institutions for their scientific impact based on bibliometric indices: numbers of publications and citations. Decisions about hiring, retention, promotion, tenure and compensation of academic scientists have been based on bibliometric indices for decades [17–30], and these readily-available quantitative bibliometric indices could easily be used to determine the impact of funded research projects.

A number of investigators have suggested that percentile scores and/or bibliometric indices are appropriate variables for determining the predictive validity of peer review [31–35], but there has not yet been a careful examination of whether retrospective studies using those variables are appropriate for the assessment of the predictive validity of peer review at NIH. Tests of the predictive validity of a screening procedure are dependent on the inclusion of all cases or a sample of cases selected at random across the full range of the population that is screened [36]. Such correlational studies are also dependent on the use of a valid criterion measure (e.g., a valid measure of scientific impact) on the same elements or cases evaluated with the screening procedure [37]. The present study was conducted to determine if retrospective analyses of funded research applications, using percentile scores from peer review as the predictor variable and bibliometric indices as criterion measures of scientific impact, satisfy the requirements for tests of predictive validity.

Methods

The R01 is the most commonly used mechanism for funding research at the NIH. It typically supports a discrete, specified, and circumscribed project of up to 5 years duration, to be performed by applicants in an area representing their specific interests and competencies, based on the mission of the NIH (see http://grants.nih.gov/grants/funding/r01.htm). In 2007 and 2008, 45,874 new R01 applications were considered for funding by the funding institutes at NIH; peer review of 87% or 39,888 of those applications were managed by the Center for Scientific Review (CSR). Of those 39,888 applications, 1.4% were incomplete or had errors or were reviewed in meetings that did not assign percentile scores. The remaining 39,337 records from the NIH IMPAC II database were included in the dataset analyzed in the present study.

Each application is reviewed in-depth by three assigned reviewers, and the average of the preliminary scores from those assigned reviewers is used to determine which applications will be discussed by the full committees in the review meetings. Usually, about half of the applications reviewed by each committee—the applications with the better average preliminary scores

—are discussed in each review meeting. For the applications discussed in the review meeting, all eligible committee members assign an overall impact score, and the average of those scores is used to calculate the percentile score. An application's percentile score is the percentage of all applications reviewed by a study section with average overall impact scores better than or equal to that application (for a more detailed description of the review and scoring procedure see [38]).

Records for awarded applications included the percentile score, the total budget requested, the total budget committed by the funding institute at the time of the award, the requested duration of the project, the approved duration of the project, and the number of days between the review meeting and the date the award was issued. This dataset includes applications assigned to all the institutes at the NIH, representing all areas of research, including applications that were discussed and not discussed, and applications that were funded and not funded.

Results

Overall, 48% of the applications were discussed (n = 19,049), and 17.4% (n = 6,830) were funded. As expected, most of the applications with the best percentile scores were funded, and fewer and fewer applications were funded as the percentile scores increased. For example, 95–96% of applications with scores in the top 10 percentile were funded, but only 86% of the applications in the 10.1–15 percentile range and only 57% of the applications in the 15.1–20 percentile range were funded (Fig 1A).

Among the funded applications, almost half had scores in the top 10 percentile, and 97% of all funded applications had scores in the top 30 percentile (Fig 1B). Only 3% of funded applications were beyond the 30th percentile, including several at greater than the 50th percentile.

Among applications in the 25.1–30 percentile range, 6 applications were reviewed for each application that was funded; for applications with percentile scores in the 30.1–35 percentile range, 16 applications were reviewed for each application funded; and for applications in the 35.1–40.0 percentile range, more than 32 applications were reviewed for each application funded (Fig 1C).

Approximately 15% of the funded applications were selected 'out of order'. In other words, 15% of the funded applications in the present study would not have been funded if the funding decision was based purely on peer review scores, applications with better peer review scores would have been funded instead. The percentage of applications funded varies widely between the different institutes at the NIH, ranging in 2007–2008 from less than 10% at some institutes to more than 30% at other institutes, and the proportion of awards made 'out of order' in terms of peer review scores, also varies widely, ranging from only 3% at the National Institutes of Aging to more than 30% at some of the smallest funding institutes (see Table 1).

In addition, analyses of variance (ANOVA) of the awarded applications showed that what was funded was different from what had been initially submitted. A 5 x 2 ANOVA on the 6,830 awarded applications including percentile scores from peer review (i.e., 0.1–10, 10.1–20, etc.) and budget (i.e., requested vs. awarded) as factors in the analysis, treating budget as a repeated measure, revealed that awarded budgets were significantly smaller than requested budgets, $F(1,6825) = 769.3$, $p < 0.0001$ (Fig 2A). Furthermore, as percentile scores increased, the difference between the requested and the awarded budgets increased. The percentile score by budget interaction was statistically significant, $F(4,6825) = 36.49$, $p < 0.0001$.

A 5 x 2 ANOVA on the 6,830 awarded applications including percentile scores from peer review (i.e., 0.1–10, 10.1–20, etc.) and project duration (i.e., proposed vs. approved) as factors in the analysis, treating project duration as a repeated measure, revealed that approved project durations were significantly shorter than proposed project durations, $F(1,6825) = 602.43$,

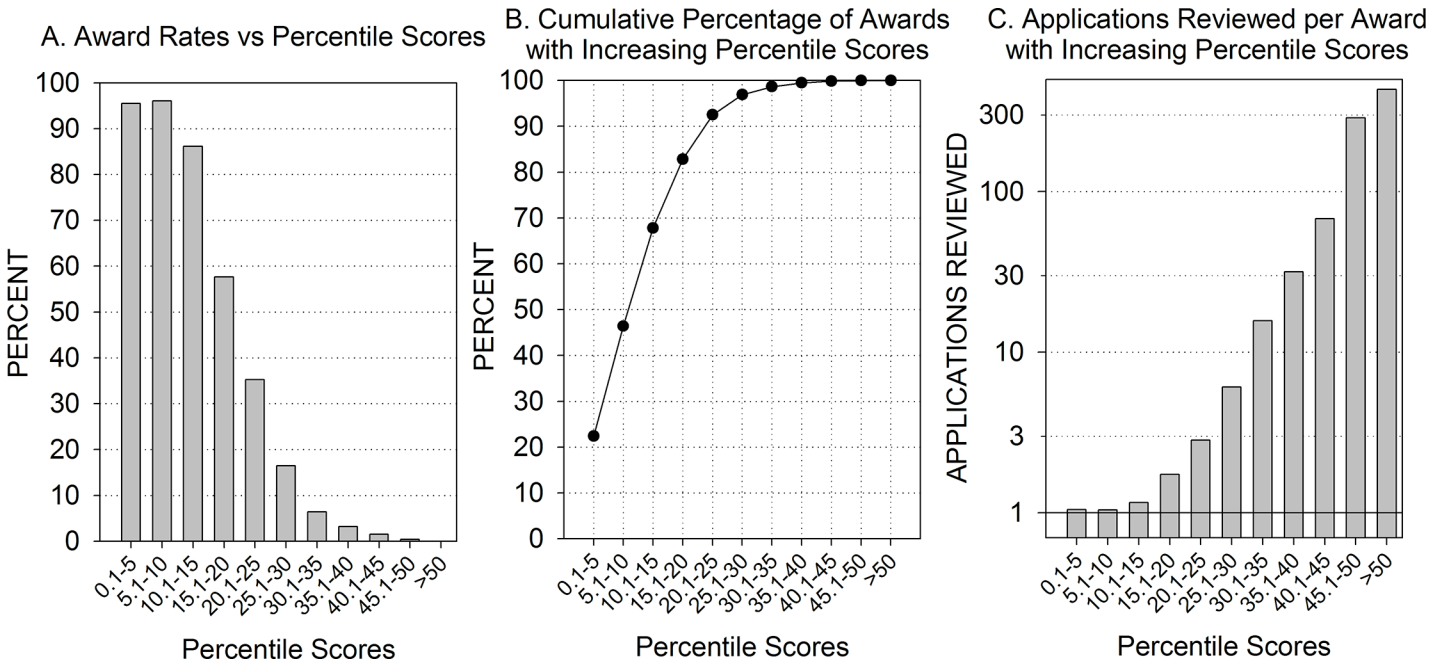

Fig 1. [A] Percent of applications funded decreases as peer review percentile scores increase. Approximately 95% of all applications with peer review percentile scores in the 0.1–10.0 range are funded, but only 3.2% of the applications with peer review scores in the 35.1–40 percentile range are funded. [B] Cumulative percentage of all funded applications with increasing peer review percentile scores. Almost 50% of all funded applications have peer review percentile scores in the 0.1–10 range, and 97% of all funded applications have peer review percentile scores equal to or less than 30. [C] Number of applications reviewed for each application funded increases as peer review percentiles increase. Almost every application reviewed with peer review percentile scores in the 0.1–10.0 range is funded, but only one of every 6 applications reviewed with peer review percentile scores in the 25.1–30.0 range is funded.

doi:10.1371/journal.pone.0126938.g001

$p < 0.0001$ (Fig 2B). As percentile scores increased, the difference between the proposed and the approved project durations increased. The percentile score by duration interaction was statistically significant, $F(4,6825) = 104.11$, $p < 0.0001$.

In addition, the delay from the review meeting until the notice of award increased as the percentile scores increased, $F(4,6825) = 283.05$, $p < 0.0001$ (Fig 2C).

Discussion

This study analyzed a large set of new (Type 1) competing R01 Research Project Grant applications reviewed at the NIH Center for Scientific Review in FY 2007 and 2008: 39,337 applications were analyzed, 48% of those applications were discussed and 17.4% were funded. Most of the applications with the best percentile scores were funded. As the percentile scores increased, fewer and fewer applications were funded and more and more applications were reviewed for each application that was funded. In addition, funded projects were different from the applications that were submitted and evaluated during peer review, and the differences between the initial peer reviewed applications and the projects that were eventually funded increased as the peer review scores increased. Furthermore, the delay from the review meeting until the notice of award increased as the percentiles increased.

The results of the present study demonstrate several reasons why retrospective studies of funded applications might fail to detect a strong linear relationship between peer review estimates of potential impact and subsequent measures of actual impact. First, tests of the predictive validity of a screening procedure are dependent on the inclusion of the full population or a

Table 1. New R01s by Funding Institute.

Institute	Institute Size (% of NIH Budget)	Award Rate 2007–2008	Awards with Percentile Scores Above Award Rates
NCI	16.4%	18%	13%
NIAID	15.1%	16%	12%
NHLBI	10.0%	19%	10%
NIGMS	6.6%	22%	17%
NIDDK	6.4%	19%	18%
NINDS	5.3%	18%	18%
NIMH	4.8%	19%	15%
NICHD	4.3%	16%	10%
NCRR	3.9%	18%	24%
NIA	3.6%	19%	3%
NIDA	3.4%	21%	14%
NIEHS	2.5%	16%	30%
NEI	2.3%	24%	22%
NIAMS	1.7%	18%	12%
NHGRI	1.7%	28%	12%
NIAAA	1.5%	24%	8%
NIDCD	1.3%	25%	21%
NIDCR	1.3%	21%	15%
NLM	1.1%	21%	20%
NIBIB	1.0%	19%	9%
NINR	0.5%	22%	30%
NCCAM	0.4%	10%	42%
FIC	0.2%	31%	21%

doi:10.1371/journal.pone.0126938.t001

sample of cases selected at random across the full range of the population that is screened. However, only a small percentage of NIH applications are funded, and those funded applications are restricted to only a portion of the entire range of applications reviewed: 97% of funded applications have peer review scores at the 30th percentile or better. Such restriction of range confounds studies of predictive validity. For example, Thorndike developed a personnel screening test to determine which applicants were more likely to successfully complete pilot training school. Scores on the screening test were correlated with successful completion of pilot training at r = 0.64. However, once training was limited to the 13% of applicants with the best scores on the screening test, the restriction of range reduced the apparent correlation between test scores and successful completion of pilot training from r = 0.64 to only r = 0.18 [36].

In addition to restriction of range, the present study underlines an NIH award process in which peer review is only the first level of review at NIH, and provides evidence of the significance of the second round of review conducted by the funding institutes which consider the initial peer review scores as only a part of their funding decisions. The funding institutes at NIH identify what they feel are the best applications that most deserve to be funded. They do this by closely examining the applications and the peer review scores and written critiques. Consistent with suggestions that at least 8% of the NIH budget should be allocated to discretionary funding of high-risk, high-reward research managed by program managers at the funding institutes [39], approximately 15% of the funded applications are selected 'out of order' by the funding institutes. In other words, 15% of the funded applications in the present study

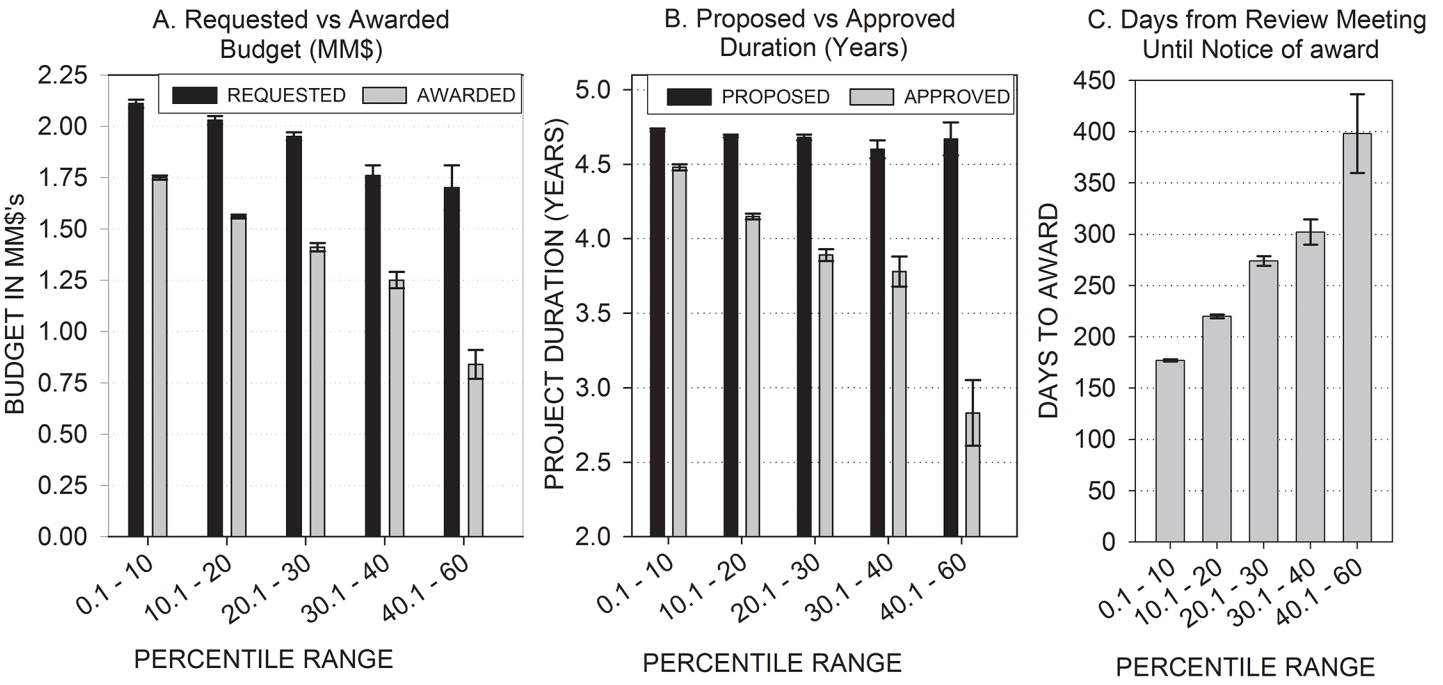

Fig 2. Awarded applications are classified by the range of percentile scores: 0.1–10 (N = 3,169), 10.1–20 (N = 2,486), 20.1–30 (N = 961), 30.1–40 (N = 178), and 40.1–60 (N = 36).

doi:10.1371/journal.pone.0126938.g002

would not have been funded if the funding decision was based purely on peer review scores, applications with better peer review scores would have been funded instead.

Especially among those applications with worse peer review scores, the funding institutes identify and 'cherry-pick' those few applications or parts of applications that they believe stand out from the rest of the applications in the same range, and they negotiate with applicants to address issues identified during peer review (see 'Negotiation of Competing Award' at http://grants.nih.gov/grants/managing_awards.htm). The extent of those negotiations and changes is reflected in the increasing differences between proposed and awarded budgets and project durations, and the increasing times from review to award. Critical information can be added, weaknesses in the design can be corrected and flawed or unnecessary experiments can be dropped.

This is not evidence of inappropriate or unethical behavior on the part of the funding institutes. They not only have the authority, they have an obligation to identify and fund the best projects. However, this means that projects are selected, revised and funded in a way that is not amenable to rigorous retrospective examination of the predictive validity of peer review scores. In addition to severe restriction of range, funded projects, especially those with worse peer review scores, are not selected at random and are not representative of other applications in the same range, and many of them are different from the projects that were initially reviewed. But the peer review scores remain unchanged: they are not revised to reflect the changes and improvements that are made during negotiations with the funding institutes. Therefore, it is not appropriate to expect that the peer review scores assigned to the original applications should be correlated with measures of the impact of the funded projects that have been revised and improved before they were eventually funded.

In addition to the issues related to peer review scores discussed above, there is also a problem with the use of citation metrics as measures of scientific impact. Citation metrics by

themselves are not accepted as valid or widely used for that purpose. Reward systems shape behavior, and the current reward system based primarily on primacy of discovery and numbers of publications and citations has shaped standards of practice in ways that facilitate achievement of those objectives [25,40–45]. For example, scientists have little incentive and rarely replicate reports from other investigators [46–49] or publish results that fail to support their own hypotheses [50–55].

Papers reporting novel, robust and statistically significant effects are more likely to be published in high-impact journals and to be more highly cited, and in order to produce novel, robust, statistically significant effects, scientists often approach their research as advocates, intent on producing and publishing results that confirm and support their hypotheses [56–60] without including adequate methodological controls to prevent their unconscious biases from affecting their results [61–65]. Scientists also often conduct a large number of small studies and use exploratory analytical techniques that virtually ensure the identification of robust, novel phenomena and hypotheses but fail to report the use of exploratory, post hoc analyses or conduct the appropriate confirmatory studies [66–71].

In addition, scientists select other papers to cite based primarily on their rhetorical utility, to persuade their readers of the value and integrity of their own work: papers are not selected for citation primarily based on their relevance or validity [72,73]. Even the father of the Science Citation Index (SCI), Eugene Garfield, noted that citations reflect the 'utility' of the source, not their scientific elegance, quality or impact [74]. Authors cite only a small fraction of relevant sources [75,76], and studies reporting robust, statistically significant results that support the author's agenda have greater utility and are cited much more often than equally relevant studies that report small or non-statistically-significant effects [75–81].

Given the emphasis on primacy of discovery and bibliometric indices, these strategies are clearly beneficial for individual scientists and do not constitute research misconduct; but, it is becoming more and more widely recognized that they produce problems of reproducibility of scientific findings and are therefore a source of waste and inefficiency [82]. Replication studies are rarely conducted [46–49], valid but small or non-statistically significant results are often not detected [83,84] or published [50–55], unnecessary studies are conducted because previous research results have not been published or cited [85,86], and uncontrolled biases lead to the publication of a large number of false positives [65,70,87–89]. In the vast majority of publications reporting novel phenomena, the effects are exaggerated or invalid [48–50,90–93], and high citation numbers do not provide assurance of the quality or the validity of the results [94,95]. Even studies reporting robust effects and cited more than 1,000 times are often not valid or reproducible [96,97].

Moreover, the evidence suggests that the magnitude of this problem is growing. With more and more highly qualified scientists, the 'publish or perish' culture continues to become ever more competitive, demanding larger and larger numbers of publications and citations in order to be hired, retained, promoted and well-compensated [98–104]. Clearly, further increasing the emphasis on the numbers of publications and citations produced by funded applications would not increase productivity, but would likely increase the problems already largely attributed to the current over-emphasis on bibliometric indices.

Instead, there is a growing consensus that scientific progress and productivity can best be increased by providing incentives to increase the integrity of the scientific literature, largely in ways that will reduce the number of publications produced and the proportion of studies reporting robust, novel effects that tend to be most highly cited [82,86,105,106]. Suggested changes to standards of practice include conducting full literature reviews and meta-analyses, citing all relevant studies, not just those that support the author's rationale; conducting power analyses and using adequate sample sizes to detect expected effects; including and clearly

communicating quality controls to prevent bias from affecting the results; clearly distinguishing between planned analyses and post-hoc exploratory analyses; and publishing results of studies in a timely manner, even if the results fail to support the investigator's hypotheses or detect any statistically significant effect. In response, the leadership at NIH has initiated changes in the NIH peer review process to incentivize quality and reproducibility over quantity [107], and one estimate suggests that such changes could significantly increase scientific productivity [82].

Conclusions

An appropriate test of the predictive validity of the peer review process has not yet been conducted. Such a test would need to include funded projects selected at random across the entire range of applications, and the projects would have to be conducted without changes or improvements based on issues identified during the peer review process. The impact of those applications would also have to be based on measures that appropriately value the impact and validity of the results. Citation numbers alone are not appropriate for that purpose, in part because citation numbers are often higher for studies that report exaggerated or invalid results [90,91,108]. Further increasing the emphasis on the numbers of publications and citations produced by funded applications, as some studies have suggested [32,34], would exacerbate the waste and inefficiencies already attributed to the current over-emphasis on bibliometric indices. Instead, the leadership at NIH has initiated changes in the NIH peer review process to incentivize quality and reproducibility over quantity [107], and one estimate suggests that such changes could significantly increase scientific productivity [82].

Acknowledgments

Disclaimer: The views expressed in this article are those of the authors and do not necessarily represent those of CSR, NIH or the US Dept. of Health and Human Services.

Author Contributions

Conceived and designed the experiments: MDL. Analyzed the data: MDL. Wrote the paper: MDL RKN.

References

1. Ehrlich E. An economic engine: NIH research, employment, and the future of the medical innovation sector. United for Medical Research. 2011.
2. Makomva K, Mahan D. In your own backyard: how NIH funding helps your state's economy. Families USA. 2008.
3. Tatsioni A, Vavva E, Ioannidis JPA. Sources of funding for Nobel Prize-winning work: Public or private? FASEB J. 2010; 24: 1335–1339. doi: 10.1096/fj.09-148239 PMID: 20056712
4. Zinner DE. Medical R&D at the turn of the millennium. Health Aff. 2001; 20: 202–209. PMID: 11558704
5. Tripp S, Grueber M. Economic impact of the Human Genome Project. Battelle Memorial Institute. 2011.
6. Narin F, Hamilton KS, Olivastro D. The increasing linkage between U.S. technology and public science. Research Policy. 1997; 26: 317–330.
7. McMillan GS, Narin F, Deeds DL. An analysis of the critical role of public science in innovation: The case of biotechnology. Research Policy. 2000; 29: 1–8.
8. Stevens AJ, Jensen JJ, Wyller K, Kilgore PC, Chatterjee S, Rohrbaugh ML. The role of public-sector research in the discovery of drugs and vaccines. N Engl J Med. 2011; 364: 535–541. doi: 10.1056/NEJMsa1008268 PMID: 21306239

9. Chatterjee SK, Rohrbaugh ML. NIH inventions translate into drugs and biologics with high public health impact. Nat Biotechnol. 2014; 32: 52–58. doi: 10.1038/nbt.2785 PMID: 24406928
10. Arias E. United states life tables, 2009. National Vital Statistics Reports. 2013; 62. PMID: 24979975
11. Manton KG, Gu X, Lamb VL. Change in chronic disability from 1982 to 2004/2005 as measured by long-term changes in function and health in the U.S. elderly population. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103: 18374–18379. PMID: 17101963
12. Mandel R. A Half Century of Peer Review, 1946–1996. Miami, FL: HardPress Publishing. 1996.
13. Kaplan D. POINT: Statistical analysis in NIH peer review—Identifying innovation. FASEB J. 2007; 21: 305–308. PMID: 17267383
14. Kirschner M. A perverted view of "impact". Science (New York, N Y). 2013; 340: 1265. doi: 10.1126/science.1240456 PMID: 23766298
15. Nicholson JM, Ioannidis JP. Research grants: Conform and be funded. Nature. 2012; 492: 34–36. 492034a [pii];doi: 10.1038/492034a PMID: 23222591
16. Demicheli V, Di Pietrantonj C. Peer review for improving the quality of grant applications. Cochrane Database of Systematic Reviews. 2007.
17. Katz DA. Faculty Salaries, Promotions, and Productivity at a Large University. The American Economic Review. 1973; 63: 469–477.
18. Salthouse TA, McKeachie WJ, Lin YG. An Experimental Investigation of Factors Affecting University Promotion Decision: A Brief Report. The Journal of Higher Education. 1978; 49: 177–183.
19. Hamermesh D, Johnson G, Weisbrod B. Scholarship, Citations and Salaries: Economic Rewards in Economics. Southern Economic Journal. 1982; 49: 472–481.
20. Sheldon PJ, Collison FM. Faculty review criteria in tourism and hospitality. Ann Tour Res. 1990; 17: 556–567.
21. Street DL, Baril CP. Scholarly accomplishments in promotion and tenure decisions of accounting faculty. J Account Educ. 1994; 12: 121–139.
22. Moore WJ, Newman RJ, Turnbull GK. Reputational capital and academic pay. Econ Inq. 2001; 39: 663–671.
23. Selpel MMO. Assessing publication for tenure. J Soc Work Educ. 2003; 39: 79–88.
24. Adler NJ, Harzing AW. When knowledge wins: Transcending the sense and nonsense of academic rankings. Acad Manage Learn Educ. 2009; 8: 72–95.
25. MacDonald S, Kam J. Aardvark, et al.: Quality journals and gamesmanship in management studies. J Inf Sci. 2007; 33: 702–717.
26. Franzoni C, Scellato G, Stephan P. Changing incentives to publish. Science (New York, N Y). 2011; 333: 702–703. doi: 10.1126/science.1197286 PMID: 21817035
27. Shao J, Shen H. The outflow of academic papers from China: Why is it happening and can it be stemmed? Learn Publ. 2011; 24: 95–97.
28. O'Keefe S, Wang TC. Publishing pays: Economists' salaries reflect productivity. Soc Sci J. 2013; 50: 45–54.
29. Fairweather JS. Beyond the Rhetoric: Trends in the Relative Value of Teaching and Research in Faculty Salaries. The Journal of Higher Education. 2005; 76: 401–422.
30. Arthur MD. What is a Citation Worth? The Journal of Human Resources. 1986; 21: 200–215.
31. Berg, JM. 6-2-2014 Productivity Metrics and Peer Review Scores [Web log post]. Available http://loop.nigms.nih.gov/2011/06/productivity-metrics-and-peer-review-scores/
32. Danthi N, Wu CO, Shi P, Lauer M. Percentile ranking and citation impact of a large cohort of national heart, lung, and blood institute-funded cardiovascular R01 grants. Circulation Research. 2014; 114: 600–606. doi: 10.1161/CIRCRESAHA.114.302656 PMID: 24406983
33. Gallo SA, Carpenter AS, Irwin D, McPartland CD, Travis J, Reynders S, et al. The validation of peer review through research impact measures and the implications for funding strategies. PLoS ONE. 2014; 9: e106474. doi: 10.1371/journal.pone.0106474 PMID: 25184367
34. Kaltman JR, Evans FJ, Danthi NS, Wu CO, DiMichele DM, Lauer MS. Prior publication productivity, grant percentile ranking, and topic-normalized citation impact of NHLBI cardiovascular R01 grants. Circ Res. 2014; 115: 617–624. doi: 10.1161/CIRCRESAHA.115.304766 PMID: 25214575
35. Scheiner SM, Bouchie LM. The predictive power of NSF reviewers and panels. Frontiers Ecol Envir. 2013; 11: 406–407.
36. Thorndike Robert L. Personnel Selection: Test and Measurement Techniques. New York, NY: John Wiley and Sons, Inc. 1949.

37. Nunnally JC. Psychometric Theory. New York: McGraw-Hill Book Company. 1978.
38. Johnson VE. Statistical analysis of the National Institutes of Health peer review system. Proc Natl Acad Sci U S A. 2008; 105: 11076–11080. doi: 10.1073/pnas.0804538105 PMID: 18663221
39. National Academy of Sciences, National Academy of Engineering and Institute of Medicine Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Future. Washington, D.C.: The National Academies. 2007.
40. Van Dalen HP, Henkens K. Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. J Am Soc Inf Sci Technol. 2012; 63: 1282–1293.
41. Lawrence PA. Lost in publication: How measurement harms science. Ethics Sci Environm Polit. 2008; 8: 9–11.
42. Lawrence PA. The politics of publication. Nature. 2003; 422: 259–261. PMID: 12646895
43. Abbott A, Cyranoski D, Jones N, Maher B, Schiermeier Q, Van Noorden R. Metrics: Do metrics matter? Nature. 2010; 465: 860–862. doi: 10.1038/465860a PMID: 20559361
44. Martinson BC, Anderson MS, De Vries R. Scientists behaving badly. Nature. 2005; 435: 737–738. PMID: 15944677
45. Anderson MS, Ronning EA, De Vries R, Martinson BC. The perverse effects of competition on scientists' work and relationships. Sci Eng Ethics. 2007; 13: 437–461. PMID: 18030595
46. Bornstein RF. Publication politics, experimenter bias and the replication process in social science research. J Soc Behav Pers. 1990; 5: 71–81.
47. Collins HM. Changing order: replication and induction in scientific practice. Chicago: University of Chicago Press. 1992.
48. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003; 33: 177–182. PMID: 12524541
49. Vineis P, Manuguerra M, Kavvoura FK, Guarrera S, Allione A, Rosa F, et al. A field synopsis on low-penetrance variants in DNA repair genes and cancer susceptibility. J Natl Cancer Inst. 2009; 101: 24–36. doi: 10.1093/jnci/djn437 PMID: 19116388
50. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483: 531–533. 483531a [pii];doi: 10.1038/483531a PMID: 22460880
51. Benatar M. Lost in translation: treatment trials in the SOD1 mouse and in human ALS. Neurobiol Dis. 2007; 26: 1–13. PMID: 17300945
52. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr. Publication bias and clinical trials. Control Clin Trials. 1987; 8: 343–353. PMID: 3442991
53. Dickersin K, Min YI. Publication bias: the problem that won't go away. Ann N Y Acad Sci. 1993; 703: 135–146. PMID: 8192291
54. Frank R, Heather MF, Susann F. Is there evidence of publication biases in JDM research? Judgment and Decision Making. 2011; 6: 870–881.
55. Su J, Li X, Cui X, Li Y, Fitz Y, Hsu L, et al. Ethyl pyruvate decreased early nuclear factor-kappaB levels but worsened survival in lipopolysaccharide-challenged mice. Crit Care Med. 2008; 36: 1059–1067. doi: 10.1097/CCM.0B013E318164403B PMID: 18176313
56. Mahoney MJ, DeMonbreun BG. Psychology of the scientist: An analysis of problem-solving bias. Cognitive Therapy and Research. 1977; 1: 229–238.
57. Mahoney MJ. Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. Cognitive Therapy and Research. 1977; 1: 161–175.
58. Mahoney MJ, Kimper TP. From ethics to logic: a survey of scientists. In: Scientist as Subject: The Psychological Imperative. New York, NY: Percheron Press. 2004. pp. 187–193.
59. Mahoney MJ. Scientist as Subject: The Psychological Imperative. Clinton Corners, NY: Percheron Press. 2004.
60. Mitroff II. The Subjective Side of Science: A Philosophical Inquiry into the Psychology of the Apollo Moon Scientists. New York, NY: Elsevier. 1974.
61. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet. 2001; 2: 91–99. PMID: 11253062
62. Colhoun HM, McKeigue PM, Davey SG. Problems of reporting genetic associations with complex outcomes. Lancet. 2003; 361: 865–872. PMID: 12642066
63. Dirnagl U. Bench to bedside: the quest for quality in experimental stroke research. J Cereb Blood Flow Metab. 2006; 26: 1465–1478. PMID: 16525413

64. van der Worp HB, Sena ES, Donnan GA, Howells DW, Macleod MR. Hypothermia in animal models of acute ischaemic stroke: a systematic review and meta-analysis. Brain. 2007; 130: 3063–3074. PMID: 17478443

65. Lindner MD. Clinical attrition due to biased preclinical assessments of potential efficacy. Pharmacol Ther. 2007; 115: 148–175. PMID: 17574680

66. Hartwig F, Dearing BE. Exploratory data analysis. Beverly Hills: Sage Publications. 1979.

67. Hoaglin DC, Mosteller F, Tukey JW. Understanding robust and exploratory data analysis. New York: John Wiley and Sons, Inc. 1983.

68. Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet. 2005; 365: 454–455. PMID: 15705441

69. Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. Control Clin Trials. 1997; 18: 530–545. PMID: 9408716

70. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science. 2011; 22: 1359–1366. doi: 10.1177/0956797611417632 PMID: 22006061

71. Kerr NL. HARKing: Hypothesizing after the results are known. Pers Soc Psychol Rev. 1998; 2: 196–217. PMID: 15647155

72. Brooks TA. Private acts and public objects: an investigation of citer motivations. Journal of the American Society of Information Science. 1985; 36: 223–229.

73. Gilbert GN. Referencing as persuasion. Social Studies of Science. 1977; 7: 113–122.

74. Garfield E. Is citation analysis a legitimate evaluation tool? Scientometrics. 1979; 1: 359–375.

75. Robinson KA, Goodman SN. A systematic examination of the citation of prior research in reports of randomized, controlled trials. Ann Intern Med. 2011; 154: 50–55. doi: 10.7326/0003-4819-154-1-201101040-00007 PMID: 21200038

76. Greenberg SA. How citation distortions create unfounded authority: Analysis of a citation network. BMJ. 2009; 339: 210–213.

77. Schrag M, Mueller C, Oyoyo U, Smith MA, Kirsch WM. Iron, zinc and copper in the Alzheimer's disease brain: a quantitative meta-analysis. Some insight on the influence of citation bias on scientific opinion. Prog Neurobiol. 2011; 94: 296–306. S0301-0082(11)00072-4 [pii];doi: 10.1016/j.pneurobio.2011.05.001 PMID: 21600264

78. Chapman S, Ragg M, McGeechan K. Citation bias in reported smoking prevalence in people with schizophrenia. Aust New Zealand J Psychiatry. 2009; 43: 277–282.

79. Jannot AS, Agoritsas T, Gayet-Ageron A, Perneger TV. Citation bias favoring statistically significant studies was present in medical research. J Clin Epidemiol. 2013; 66: 296–301. doi: 10.1016/j.jclinepi.2012.09.015 PMID: 23347853

80. Kjaergard LL, Gluud C. Citation bias of hepato-biliary randomized clinical trials. J Clin Epidemiol. 2002; 55: 407–410. PMID: 11927210

81. Gotzsche PC. Reference bias in reports of drug trials. Br Med J (Clin Res Ed). 1987; 295: 654–656. PMID: 3117277

82. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet. 2009; 374: 86–89.

83. Bakker M, van Dijk A, Wicherts JM. The Rules of the Game Called Psychological Science. Perspectives on Psychological Science. 2012; 7: 543–554.

84. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14: 365–376. nrn3475 [pii];doi: 10.1038/nrn3475 PMID: 23571845

85. Chapman SJ, Shelton B, Mahmood H, Fitzgerald JE, Harrison EM, Bhangu A. Discontinuation and non-publication of surgical randomised controlled trials: Observational study. BMJ (Online). 2014; 349.

86. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Lancet. 2014; 383: 156–165. doi: 10.1016/S0140-6736(13)62229-1 PMID: 24411644

87. Counsell CE, Clarke MJ, Slattery J, Sandercock PA. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? BMJ. 1994; 309: 1677–1681. PMID: 7819982

88. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet. 2001; 29: 306–309. PMID: 11600885

89. Munafo MR, Stothart G, Flint J. Bias in genetic association studies and impact factor. Mol Psychiatry. 2009; 14: 119–120. doi: 10.1038/mp.2008.77 PMID: 19156153

90. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005; 2: e124. PMID: 16060722

91. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008; 19: 640–648. doi: 10.1097/EDE.0b013e31818131e7 PMID: 18633328

92. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011; 10: 712. nrd3439-c1 [pii];doi: 10.1038/nrd3439-c1 PMID: 21892149

93. Steward O, Popovich PG, Dietrich WD, Kleitman N. Replication and reproducibility in spinal cord injury research. Exp Neurol. 2012; 233: 597–605. S0014-4886(11)00239-1 [pii];doi: 10.1016/j.expneurol.2011.06.017 PMID: 22078756

94. Reinstein A, Hasselback JR, Riley ME, Sinason DH. Pitfalls of using citation indices for making academic accounting promotion, tenure, teaching load, and merit pay decisions. Issues Account Educ. 2011; 26: 99–131.

95. Browman HI, Stergiou KI. Factors and indices are one thing, deciding who is scholarly, why they are scholarly, and the relative value of their scholarship is something else entirely. Ethics Sci Environm Polit. 2008; 8: 1–3.

96. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005; 294: 218–228. PMID: 16014596

97. Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG, Ioannidis JP. Establishment of genetic associations for complex diseases is independent of early study findings. Eur J Hum Genet. 2004; 12: 762–769. PMID: 15213707

98. Moore WJ, Newman RJ, Turnbull GK. Do academic salaries decline with seniority? J Labor Econ. 1998; 16: 352–366.

99. Balogun JA, Sloan PE, Germain M. Core values and evaluation processes associated with academic tenure. Percept Mot Skills. 2007; 104: 1107–1115. PMID: 17879644

100. Youn TIK, Price TM. Learning from the experience of others: The evolution of faculty tenure and promotion rules in comprehensive institutions. J High Educ. 2009; 80: 204–237.

101. Graber M, Launov A, Wälde K. Publish or perish? The increasing importance of publications for prospective economics professors in Austria, Germany and Switzerland. Ger Econ Rev. 2008; 9: 457–472.

102. Pilcher ES, Kilpatrick AO, Segars J. An assessment of promotion and tenure requirements at dental schools. J Dent Educ. 2009; 73: 375–382. PMID: 19289726

103. Fanelli D. Do pressures to publish increase scientists' bias? An empirical support from US states data. PLoS ONE. 2010; 5.

104. Fanelli D. Negative results are disappearing from most disciplines and countries. Scientometrics. 2012; 90: 891–904.

105. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014; 383: 166–175. doi: 10.1016/S0140-6736(13)62227-8 PMID: 24411645

106. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, et al. Biomedical research: Increasing value, reducing waste. The Lancet. 2014; 383: 101–104. doi: 10.1016/S0140-6736(13)62329-6 PMID: 24411643

107. Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature. 2014; 505: 612–613. PMID: 24482835

108. Young NS, Ioannidis JP, Al-Ubaydli O. Why current publication practices may distort science. PLoS Med. 2008; 5: e201. doi: 10.1371/journal.pmed.0050201 PMID: 18844432