frontiers in Research Metrics and Analytics

REVIEW
published: 23 August 2018
doi: 10.3389/frma.2018.00022

External Tests of Peer Review Validity Via Impact Measures

Stephen A. Gallo and Scott R. Glisson*

American Institute of Biological Sciences, McLean, VA, United States

Peer review is used commonly across science as a tool to evaluate the merit and potential impact of research projects and make funding recommendations. However, potential impact is likely to be difficult to assess ex-ante; some attempts have been made to assess the predictive accuracy of these review decisions using impact measures of the results of the completed projects. Although many outputs, and thus potential measures of impact, exist for research projects, the overwhelming majority of evaluation of research output is focused on bibliometrics. We review the multiple types of potential impact measures with an interest in their application to validate review decisions. A review of the current literature on validating peer review decisions with research output impact measures is presented here; only 48 studies were identified, about half of which were US based and sample size per study varied greatly. 69% of the studies employed bibliometrics as a research output. While 52% of the studies employed alternative measures (like patents and technology licensing, post-project peer review, international collaboration, future funding success, securing tenure track positions, and career satisfaction), only 25% of all projects used more than one measure of research output. Overall, 91% of studies with unfunded controls and 71% of studies without such controls provided evidence for at least some level of predictive validity of review decisions. However, several studies reported observing sizable type I and II errors as well. Moreover, many of the observed effects were small and several studies suggest a coarse power to discriminate poor proposals from better ones, but not amongst the top tier proposals or applicants (although discriminatory ability depended on the impact metric). This is of particular concern in an era of low funding success, where many top tier proposals are unfunded. More research is needed, particularly in integrating multiple types of impact indicators in these validity tests, as well as considering the context of the research outputs relative to goals of the research program and concerns for reproducibility, translatability and publication bias. In parallel, more research is needed focusing on the internal validity of review decision making procedures and reviewer bias.

Keywords: research funding, peer review, impact metrics, validity, grant

OPEN ACCESS

Edited by:
George Chacko,
NET eSolutions Corporation (NETE),
United States

Reviewed by:
Peter Van Den Besselaar,
VU University Amsterdam,
Netherlands
Kevin Boyack,
SciTech Strategies, Inc., United States

****Correspondence:***
Stephen A. Gallo
sgallo@aibs.org

Received: *11 April 2018*
Accepted: *20 July 2018*
Published: *23 August 2018*

Citation:
Gallo SA and Glisson SR (2018) External Tests of Peer Review Validity Via Impact Measures. Front. Res. Metr. Anal. 3:22. doi: 10.3389/frma.2018.00022

GRANT PEER REVIEW AND IMPACT ASSESSMENT

Most would generally agree the purpose of biomedical research is to advance knowledge for societal benefit, with the hope of favorably impacting disease outcomes and improving global health. Indeed, the National Institutes of Health (NIH), the world's largest funder of biomedical research, characterizes their mission as to "seek fundamental knowledge about the nature and

behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability" (NIH, 2017). To help select which research projects to fund to achieve this goal, NIH and other funders rely on a peer review process to assess the quality of the research approach and methodologies proposed, the feasibility of the investigators successfully conducting the project in the proposed environment, and the level of innovation and potential significance of the project (NIH, 2014). Of these criteria, it is likely the most difficult to accurately assess is the potential significance; particularly "if the aims of the project are achieved, how will scientific knowledge, technical capability, and/or clinical practice be improved" and "how will successful completion of the aims change the concepts, methods, technologies, treatments, services, or preventative interventions that drive this field?" (NIH, 2016).

In no small part, this is due to the role of serendipity in science, which has been identified as an important component in scientific discovery (Ban, 2006; Merton and Barber, 2011; Editorial, 2018), as well as a variety of unforeseen factors which may prevent the success of a research project. Thus, even in the best of cases, the potential impact of a research project may be difficult to gauge. However, there are also reports that the decision-making process can be hampered by subjectivity and the presence of biases (Marsh et al., 2008; Ginther et al., 2011; Lee et al., 2013; Boudreau et al., 2016; Kaatz et al., 2016). As one of the chief goals of peer review is to select projects for funding of the highest scientific quality that are likely to have the greatest impact, it stands to reason that objective measurements of the actual impact of fully funded and completed projects could be assessed ex-post funding and compared to peer review evaluations, so that we may determine the predictive validity of these decisions. Similarly, objective indicators of proposal quality (e.g., track record of the applicant) could be assessed ex-ante to funding to be compared to review decisions. These external tests of validity, which compare scientific inputs and outputs to review evaluations, likely offer an important assessment of the effectiveness of review decisions in choosing the best science, although admittedly do not necessarily validate other expectations of peer review, like impartiality (Wood and Wessely, 2003).

However, a central question in scientometrics is how best to evaluate research, as many metrics have considerable limitations or are influenced by a variety of factors that are not associated with research quality or scientific impact (Nieminen et al., 2006; Bornmann et al., 2008a; Hagen, 2008; Leydesdorff et al., 2016). For instance, citation levels are influenced by the number of co-authors, journal prestige and even by whether the results are positive or negative (Callaham et al., 2002; Dwan et al., 2008; Ioannidis, 2008). Moreover, for biomedical research, the societal impact of a study is not only measured in its contribution to the knowledge base (Bornmann, 2017), but also in actual improvements to human health; however, linking the influence of individual works to the development of new therapeutics is problematic, as they rely on large bodies of work through their evolution from bench to bedside (Keserci et al., 2017). Nevertheless, as the recent Leiden manifesto points out, performance measurements should "take into account the wider socio-economic and cultural contexts," and that the "best practice uses multiple indicators to provide a more robust and pluralistic picture" (Hicks et al., 2015).

Thus, it seems a variety of impact measures should potentially be used to validate review decisions. However, at this time there has been no comprehensive review of studies in the literature, across a variety of impact measures, that have attempted to validate peer review decisions. We will review here many of these measures below, examining what has been done with respect to peer review of research funding applications, what measures still need to be explored, and what has been done to integrate these measures to achieve a more well-rounded assessment of research success and failures. It should be noted that this literature review is focused on work that is application based. That is, it includes studies that examine the ranking and funding fate of applications and applicants relative to either the quality of the input or the impact of the output from those applications and applicants after the funding decision across a variety of measures (**Figure 1**). Again, this includes only measures of external validity (external scientific quality measures for outputs and inputs) and not the internal validity of review procedures (e.g., bias, inter-rater reliability), which is beyond the scope of this review. It is based on the knowledgeable selection of relevant publications which includes both peer reviewed and non-peer reviewed articles, as some of this work has been conducted by funding agencies and published in non-traditional forums.

PUBLICATION PRODUCTIVITY AND CITATION IMPACT

The most studied research outputs are bibliometric in nature, surrounding the number of published manuscripts, the impact of the journals they were published in, the raw and normalized citation levels of these manuscripts (normalized for time and research field), the h-indices of applicants and number of manuscripts in the top 10% of all other cited papers on the topic as well as citations and papers per dollar spent (Mavis and Katz, 2003; Van Noorden, 2010; Danthi et al., 2014; Li and Agha, 2015). As mentioned above, there are limitations to bibliometric indicators due to their complex nature and may not always reflect long term impact (Wang et al., 2013). Nevertheless, this is where much of the effort to study the validation of peer review has focused. Several types of similarly structured studies have resulted, which are summarized below.

Ex Ante Impact of Applicants (Funded vs. Unfunded or Review Score)

In the last few years, several attempts have been made to examine the number of publications and their citation impact from funded and unfunded applicants. Several studies have tracked individual applicant ex ante performance before funding decisions to determine if reviewers can pick applicants with superior prior publication and citation performance. This is a powerful strategy as you can directly compare funded and unfunded applicants, and do not have to consider the effect of funding as a confounding factor on performance. Most studies show that overall funded

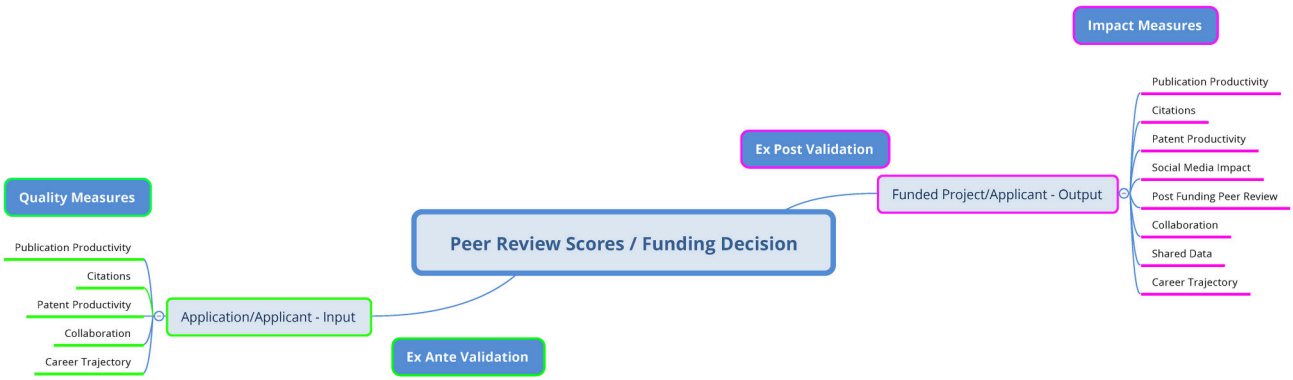

FIGURE 1 | Validation Model. Model of external validation of peer review decisions using ex ante quality measures and ex post impact metrics.

applicants outperform unfunded (Bornmann and Daniel, 2006; van den Besselaar and Leydesdorff, 2009; Bornmann et al., 2010; van Leeuwen and Moed, 2012; Cabezas-Clavijo et al., 2013) and a few studies do not (Hornbostel et al., 2009; Neufeld et al., 2013; Saygitov, 2014), although typically the differences are small and dependent on the general quality level of applicants (if all applicants are very productive, smaller differences will be observed). A couple of studies examined the ex-ante productivity of applicants relative to review scores, and found significant correlations, as well as significant biases (Wenneras and Wold, 1997; Sandstrom and Hallsten, 2008). Also, some studies show when you compare the best of unfunded applicants with funded ex ante, they are comparable (van den Besselaar and Leydesdorff, 2009; Bornmann et al., 2010; Neufeld et al., 2013), suggesting significant type II error. Some of these studies have been summarized well by Boyack et al. (2018) as well as Van den Besselaar and Sandstrom (2015). Thus, these results may suggest that while peer review may be efficient at coarse discrimination between bad and good applicants, it may be limited in its ability for fine discrimination between good and excellent applicants. However, only looking at ex-ante results makes no comment on how applicants actually perform in the future, which is what reviewers are predicting via their score, therefore it is important to make ex-post observations as well.

Ex Post Impact of Applicant and Project (Funded vs. Unfunded)

Some studies examine the productivity of funded applicants ex post in comparison to unfunded, to see if reviewers chose applicants that in the end were productive. Multiple studies show that funded applicants are at least modestly more productive and more frequently cited after the award as compared to unfunded (Armstrong et al., 1997; Mavis and Katz, 2003; Mahoney et al., 2007; Bornmann et al., 2008b, 2010; Pion and Cordray, 2008; Reinhart, 2009; Campbell et al., 2010; Jacob and Lefgren, 2011a,b; Langfeldt et al., 2012; Robitaille et al., 2015; Van den Besselaar and Sandstrom, 2015; Gush et al., 2017), although some do not (Saygitov, 2014). Interpretation of these results is difficult because it is challenging to dissociate the productivity effect of funding from the validity of the review decision. However, while general research funding is related to scientific productivity and knowledge production (Lauer, 2015; Rosenbloom et al., 2015) and papers with funding acknowledgments are linked to higher citation counts (Gok et al., 2016), the effect of specific funding on an individual's productivity is not clear; some research looking at ex ante and ex post bibliographic levels for funded applicants show no effect of funding at all (Langfeldt et al., 2012; Robitaille et al., 2015), although it seems the length of time used to capture ex post bibliometric data is an important factor (Van den Besselaar and Sandstrom, 2015). Once again, many of these studies show significant type II errors (where unfunded applicants perform well) and sometimes only limited or no differences are found between funded and unfunded applicants with similar review scores or performance (Bornmann et al., 2008b, 2010; Pion and Cordray, 2008; Jacob and Lefgren, 2011a; Van den Besselaar and Sandstrom, 2015; Gush et al., 2017) although some similar comparisons do find differences (Robitaille et al., 2015).

These ex post studies are related to the above ex ante results in that some literature has indicated that one of the strongest predictors of future citation performance is prior citation performance (Kaltman et al., 2014; Hutchins et al., 2016). Thus again, if peer review selects for applicants with higher previous productivity, it stands to reason that their post-funding productivity will be higher than unfunded applicants as well. While this could be interpreted as further validation of the peer review process, the assumption is that some investigators are simply more inherently productive than others. However, this could also be interpreted as the Matthew effect, where the rich get richer; the subset with access to research funding have more opportunities to be productive, which leads to more funding, more security and prestige, and therefore better bibliometric output (Merton, 1968; Azoulay et al., 2013), although some studies find no evidence of this (Boyack et al., 2018). In addition, many grant proposals are judged around the assessment of a research idea and its methodological implementation, not just the investigator's track record. Thus, it is unclear looking at an individual's career productivity alone may be an appropriate measure of success to validate review decisions; analysis of ex post productivity of individual projects is also required.

Ex Post Impact of Funded Project vs. Review Score (No Unfunded Control)

Similar studies have been performed with projects, although admittedly these are harder to conduct as productivity and impact data from unfunded projects is impossible to access or difficult to interpret. Largely what has been done is to analyze the relative confidence in funding decisions (peer review scores) of funded projects and how these relate to citation impact. One issue has been how results are normalized and computed. For instance, several studies of NIH NHLBI data calculated output results on a per dollar spent basis, as some research (Berg, 2011; Fortin and Currie, 2013; Gallo et al., 2014) has predicted diminishing returns with larger investments; these studies found no correlation between review scores and output (Danthi et al., 2014; Doyle et al., 2015). However, a large NIH study of unnormalized bibliometric data found a moderate correlation (Li and Agha, 2015). In fact, several other studies using normalized and unnormalized citation impact measures also suggested a moderate correlation (Berg, 2011; Gallo et al., 2014). When NIH data were reanalyzed without using budget normalized citation impact, a moderate correlation was observed (Lauer et al., 2015). A few other studies have found no correlation between scores and citation impact, although one was a very small sample (Scheiner and Bouchie, 2013) and the other was from the second round of review, so the level of quality across these projects was already very high (Gush et al., 2017). In fact, similar results were found with NIH data (same data set as used by Li and Agha, 2015); if the poorer scoring applications were removed from the analysis to reflect current funding rates, correlations between output and review scores disappeared (Fang et al., 2016). Again, this suggests the coarse discrimination of peer review in separating good projects from poor ones, but not good from great.

One constant in all of these analyses is the high degree of variability in grant output and impact across projects. This variability reflects the complicated and potentially biased nature of bibliometrics (e.g., dependencies on field, number of authors, and on the research results themselves), but also the role of serendipity in science (not every discovery receives the same reception, and while breakthrough discoveries are rare, they often stand on the shoulders of previous less-cited research). Many attempts to normalize citation counts for confounding factors have been made (h-index, Hirsch, 2005, m-index, Bornmann et al., 2008c; RCR, Hutchins et al., 2016) but each method has strengths and weaknesses (Van Noorden, 2010). Complicating this is the observation that reviewers may treat higher risk projects differently than straightforward ones (Boudreau et al., 2016). Given this bibliometric complexity and the inherent riskiness of research projects, a strong correlation between peer review scores and citation patterns, where better scores predict high performance projects may be unattainable. In fact, some groups have asserted that "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" (Lindner and Nakamura, 2015), as citation values many times are higher for "exaggerated or invalid results" and that papers are often selected for citation based on their "rhetorical utility" and not "primarily based on their relevance or validity."

Type I/ II Error and Peer Review Scores

While it may not be clear how to define relative success of productive projects, it is easily achievable to determine which projects published anything at all. To date there has not been an exploration of the relationship between peer review scores of projects and the likelihood of unproductive grants (funded projects yielding no publication output), despite suggestions that the failure is an important aspect to the scientific process of discovery (Firestein, 2015). To address this issue, we have re-analyzed previously published data focusing on the frequency of non-producing grants and its relationship to score. The data used in this analysis came from independent peer reviews of 227 R01 style awards (4 year, $1 million awards) funded from an anonymous biomedical research program (Gallo et al., 2014). We define type I error (ratio of unproductive grants/all grants of a given score) as projects that are funded but ultimately yield no publications after funding is completed and the grant is closed. Projects are rated on a scientific merit (SM) scale of 1–5 (1 being most meritorious). In **Figure 2** below, we observe a moderate level of correlation between the proportion of funded projects with zero resultant publications and peer review score ($R^2 = 0.23$; $p = 0.07$), with better scoring grants yielding lower error rates than poorer scoring grants (removal of the outlier at 1.2 yields an $R^2 = 0.58$; $p = 0.002$). Across the entire scoring range, the overall type I rate was 33%, with unproductive grants having a median score of 1.9 ± 0.05, vs. 1.7 ± 0.03 for productive grants (non-zero). Others have defined type I errors as lower than median performance for funded projects (using metrics like the h-index) and have estimated these values at 26–37% (Bornmann et al., 2008b), which is similar to that observed here, albeit using a less generous cut-off. The fact that nearly a third of grants were unproductive and yet 50% of those unproductive grants scored a 1.9 or better perhaps speaks not only to the level of quality but also to the level of risk involved in research projects, and that flaws which impact the score of an application may also increase the risk of unproductive projects. Indeed, some studies have suggested more novel (but potentially higher risk) applications are penalized in review score (Boudreau et al., 2016).

The rate of false negatives, or type II error, could be defined as unfunded projects that were eventually completed and were highly productive. This is clearly a more difficult aspect to measure, as there are few follow-up data linking unfunded applications and their ideas to post-review publications. As such, few studies exist assessing type II error, although some attempts have been made tracking the h-indices of successful and unsuccessful applicants, estimating type II rates as 32–48% (Bornmann et al., 2008b). Type II errors are probably highly dependent on funding success rates. While it has been shown that reviewers agree more about what should not be funded than what should (Cole and Simon, 1981), it is likely that as scores approach the funding line, there will be higher levels of type II error, which

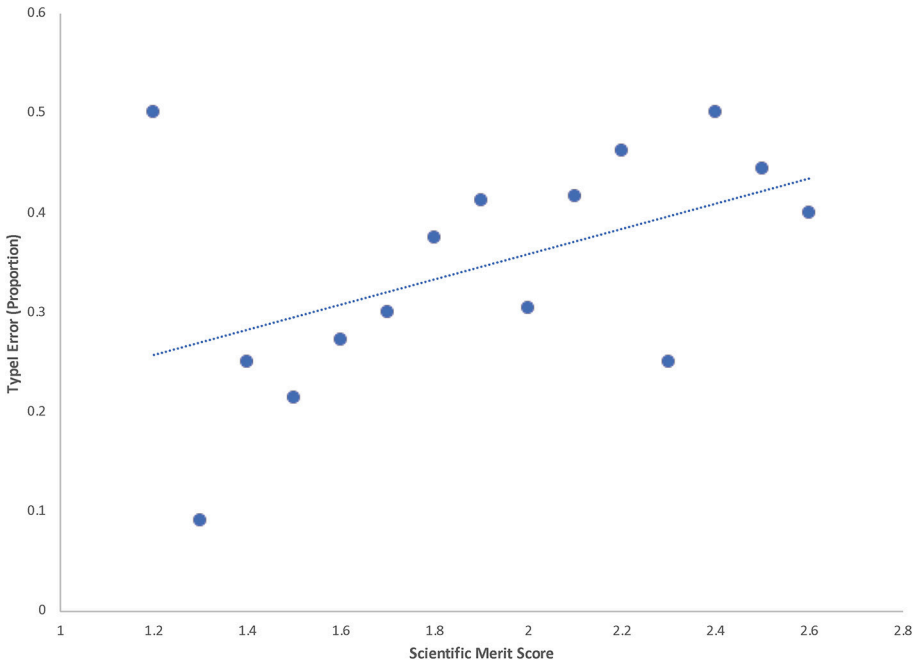

FIGURE 2 | Type I Error Rates vs. Review Score. Proportion of projects with zero publications ex-post vs. ex-ante peer review score (scale of 1–5 where 1 is the highest merit).

may result in a graph similar to **Figure 2**, although there are no such studies in the literature currently.

SOCIAL MEDIA IMPACT (ALTMETRICS)

Most publishers now enable the use of altmetrics to capture the number of tweets and other social media posts about articles, as well as capture download rates and page views. These dynamic metrics capture in real time another sense of impact, "quantifying the interest and debate an article generates, from the moment it is published (Warren et al., 2017)." While critics have mentioned that altmetrics are not yet validated and represent popularity, not necessarily impact, proponents suggest social media discussions represent a new, broader channel of communication that could reach beyond discipline and even increase engagement outside the scientific community (Sugimoto et al., 2017). Altmetrics have the capability to capture and quantify types of outputs that are missed by traditional bibliometrics. For instance, white papers and non-peer reviewed publications do not necessarily yield citations in Web of Science, but yet may be of great importance and influence on science policy. In addition, blogs, conference presentations and other alternate publications may be the only route to announce negative results, which may be unpublishable in traditional journals but are still useful and important products of the research. Although one study suggests funded research is viewed online more often than unfunded research (Didegah et al., 2017), and another has examined the relation between views, Twitter counts and post-publication peer review results of manuscripts (Bornmann, 2017), there are currently no studies in the literature directly looking at funding decisions and altmetrics (Dinsmore et al., 2014).

COLLABORATION-FOSTERING OF RESEARCH TEAMS

There is an argument to be made that high degrees of collaborations between scientists (especially interdisciplinary collaborations) addressing a common research objective yield higher creativity and innovation, as well as higher translatability (Carayol and Thi, 2005). Also, higher collaboration may enhance reproducibility (Munafo et al., 2017). Thus, tracking the actual level of collaboration (both that contained in the original proposal as well as ex post published co-authorships) may be important, especially if this is one of the goals of the research funding program. In fact, it has been shown that receiving more funding may be a result of increased collaboration (Ebadi and Schiffauerova, 2015) and may result in larger future collaborations (Adams et al., 2005). While research into collaborative scientific activities is extensive, only a few studies have looked at this directly with regard to peer review decisions; both Melin and Danell (2006) and Langfeldt et al. (2012) found successful applicants have a higher degree of ex-post international co-authorship than unsuccessful applicants and both El-Sawi et al. (2009) and Ubfal and Maffioli (2011) have found increased levels of collaboration amongst funded groups. However, Robitaille et al. (2015) found funded applicants had lower levels of ex post interdisciplinarity and Bromham et al. (2016) also notes that projects with greater interdisciplinarity

have lower funding success, even for projects with high degrees of collaboration. This may be due to the risk that interdisciplinarity brings, as some results have shown increased novelty (presumably high risk) is penalized by reviewers (Boudreau et al., 2016). This small amount of data suggests perhaps that peer review decisions can validly select projects that yield high degrees of collaboration but are not necessarily promotional of interdisciplinary research, although it also seems clear much more work needs to be done on this subject.

POST-FUNDING REVIEW OF OUTCOMES

A few studies included in this review have looked at peer review evaluation of post-funding performance and quality (Claveria et al., 2000; Mutz et al., 2015) and compared it to the ex-ante evaluation of proposals; both of these findings observed significant predictive validity of the review decisions (although the work of Mutz relies strongly on some methodological assumptions and may not represent an independent observation). Post-funding evaluations of productivity and impact likely take into account contextual factors of the research that are not represented in bibliometric numbers. The obvious downside is that conducting post-funding review panels is likely cost prohibitive, preventing its regular use. Post-publication peer review (PPPR) sites like PubPeer and F1000 may also be used to get a sense of trustworthiness and robustness of individual publications via the comments and ratings (Knoepfler, 2015). However, while one could conceivably achieve a high number of reviewers per publication and therefore a high degree of confidence in the results, there is concern for potentially low and inconsistent levels of engagement and for some reviewers, the lack of anonymity will be an issue (Dolgin, 2018). Administrative review post-funding can also be done at the funding agency level to at least determine whether a variety of non-bibliometric outcomes were achieved, which can include whether the work was finished or left incomplete, whether the stated goals were achieved, whether the results or products were disseminated (including through non-traditional pathways) and tracking the level of reproducibility of the results. One recent example of this is by Decullier et al. (2014), who found that clinical projects chosen to be funded by an agency were much more likely to be initiated than unfunded projects. However, once a project was initiated, the authors observed that the likelihood of completion was unaffected by funding status, as was whether publications would result, the timeline to publications and the number of publications. Therefore, straight interpretation of publication output may mask type II error, as the productivity level of unfunded but initiated projects was similar to that of funded ones. Thus, these types of measures provide crucial context to the interpretation of the results.

PATENTS/TECHNOLOGY DEVELOPMENT

Patents have been used as indicators of research impact, although some studies find that only about 10% of NIH grants over the last 3 decades directly yielded a patent as a product and only about 30% have work which is cited in a patent (Li et al., 2017). Other studies have shown that, to bring 5 patented therapeutics through testing and to the market required more than 100,000 papers, and nearly 20,000 NIH grants (Keserci et al., 2017; other funding sources not considered). In addition, some have argued that linkages between patents and the literature should not only rely on direct citation linkages, but on mapping analysis of whole bodies of work surrounding a concept to determine the influence of an individual (Gurney et al., 2014), further complicating analysis. Thus, attributing an individual grant to the creation and subsequent impact of a patent may be difficult, as not only do multiple research inputs cumulatively produce a patent, the success rate for producing an actual therapeutic in the market is very low (Stevens and Burley, 1997).

Nevertheless, some research has been conducted observing the predictive association of peer review scores of funded grant applications and patent production (Li and Agha, 2015); finding a decrease in score of one standard-deviation yielding 14% fewer patents. Galbraith et al. (2006) also compared peer review scores of individual funded projects to their ultimate success utilizing two metrics: (1) cooperative research and development agreements (CRADA) or licenses that were signed, SBIR or equity funding that was obtained or a product that was launched; and (2) the assessment of a senior project manager (not an author) of each technology as successful (evaluated one and a half to 3 years after the initial peer review evaluation). Using 69 early to mid-stage homeland defense technologies funded by the US DoD Center for Commercialization of Advanced Technologies (CCAT), the authors found that reviewer scores were weakly predictive of commercial success of funded projects. However, Melin and Danell (2006) found that, for a Swedish research funding program aiming to develop research with industrial applications with large, 6-year grants, funded applicants generated more patents and more spin-off companies than unfunded applicants, although the sample is small with large variation in patent output (which may in part be due to the wide breadth of scientific fields). Chai and Shih (2016) also found that firms funded by an academic-industry partnership received significantly more patents than unfunded applicant firms, although the effects depended on the size and age of the firm. These results suggest some level of review validity, although it is still unclear how and to what extent the funding can promote patent creation. It may be the direct effect is small; while some have observed small positive impacts on patent generation (Payne and Siow, 2003) or on patent originality and impact (Huang et al., 2006; Guerzoni et al., 2014), some have found no effect or even a negative effect (Sanyal, 2003; Beaudry and Kananian, 2013). Thus, patent productivity has some promise for use in tests of review validity, however future studies will likely require more subtle, nuanced approaches.

DATA SHARING

An important output of research is sharable data sets, which some have suggested have "vast potential for scientific progress" by facilitating reproducibility and allowing new questions to

be asked with old data sets (Fecher et al., 2015). In fact, data sharing is associated in some cases with increased citations rates (Piwowar et al., 2007). Yet, several studies have indicated the majority of researchers do not share their data, in part because of the lack of incentives (Tenopir et al., 2011; Fecher et al., 2015; Van Tuyl and Whitmire, 2016). Multiple platforms are available to share data through journal publication sites (e.g., PloS One) or even sites hosting unpublished manuscripts and data (e.g., Figshare). Various metrics, such as download rates or even citations of data usage can be used to potentially capture impact. Yet, while one study examined data management plans for funded and unfunded National Science Foundation (NSF) proposals and found no significant differences in plans to share data (Mischo et al., 2014), currently no studies have explored ex post data sharing and its relationship to peer review decisions.

CAREER TRACKING

Some have focused efforts on assessing impact of early career funding through tracking of PI careers, using ex-post NIH funding as a metric. One study of the Howard Hughes Medical Institute's (HHMI) research training programs for medical students found that funding through their program was associated with significantly increased levels of NIH post-doctoral funding success post-HHMI award (21%) as compared to a control group of unfunded HHMI applicants (13%; Fang and Meyer, 2003). It should be noted that funded applicants still had higher success than unfunded applicants despite similar ex-ante qualifications. In addition, when ex-ante peer review results were taken into account, similar results were also seen with the Doris Duke Charitable Foundation (DDCF) Clinical Scientist Development Award (CSDA), where a greater proportion of CSDA funded applicants received at least one R01 grant (62%) vs. highly ranked but unfunded CSDA applicants (42%; Escobar-Alvarez and Myers, 2013). Moreover, NIH itself has observed differences between similarly scored funded and unfunded K grant applicants and their relative success in acquiring additional NIH funding (56% for K grant awardees vs. 43% for unfunded; Mason et al., 2013). Similar results were found between similarly scored funded and unfunded applicants by Tesauro et al. (2013). Mavis and Katz (2003) also observed higher post-award funding rates for successful applicants compared to unsuccessful ones, although there was no control for review score. Similarly, others have shown that, despite similar qualifications, funded applicants are more successful in gaining future funding and securing tenure track positions compared to unfunded applicants (Bol et al., 2018; Heggeness et al., 2018). However, many of these observations may be the result of the funding itself enabling future funding, as well as lowered levels of resubmissions by unfunded applicants. If possible, the effect of funding itself needs to be addressed in these tests, possibly by utilizing review scores to compare the amount of funded applicant's ex post funding success, although no such studies have been done.

Other metrics along the same vein have been used as well, including career satisfaction and faculty positions attained, both of which have been observed to be higher among funded applicants compared to similarly high ex-ante performing unfunded applicants (Hornbostel et al., 2009; Bloch et al., 2014; Van den Besselaar and Sandstrom, 2015). However, while Pion and Ionescu-Pioggia (2003) also found funded applicants of the Burroughs Welcome Career Award were more successful than unfunded in securing faculty positions and in acquiring future NIH funding (Pion and Cordray, 2008), these effects were diminished when adjusted for the ex-ante qualifications of the applicants. Career satisfaction is another variable to be tracked, although only two studies have examined this (Hornbostel et al., 2009; Langfeldt et al., 2012), tracking satisfaction via survey. While these groups found higher levels of satisfaction associated with funded applicants, there were no ex-ante controls for this measure and may be a result of the funding itself. Similarly, while (Langfeldt et al., 2012) has also monitored the number of successful graduate theses created stemming from funded applicants, again this work lacks the appropriate control to address peer review decisions. On the whole, while many of these results contrast bibliometric results above (given the high level of discrimination between competitive applicants), it is clear that future studies need to de-couple the effects of funding itself from the review decision before this measure can truly test review validity.

INTEGRATION OF MULTIPLE IMPACT METRICS

Including a panel of indicators is likely to give a clearer picture of impact (Hicks et al., 2015), but they still need to be interpreted in the qualitative context of the science and the funding program (Chen, 2016), and the "right balance between comprehensiveness and feasibility must be struck" when determining how many and what type of indicators to include (Milat et al., 2015). In addition, just as reviewers weigh the relative importance of review criteria, how one weighs the importance of each indicator into the overall picture of impact is of crucial importance (Lee, 2015; Milat et al., 2015). Thus, integration of this information within a specific research context is crucial to getting an accurate picture of impact, but this is still represents a largely unexplored area, particularly with regard to validating peer review. One example of the use of multiple indicators in our survey was by Melin and Danell (2006), who found that subsequent to funding, while the number of publications was no different, funded applicants published in higher quality journals, as well as received more external funding for their group, produced more spin-off companies and produced more patents. Similarly, Hornbostel et al. (2009) found minor differences in bibliometric impact and output between funded and unfunded groups, yet both career satisfaction and number of faculty positions gained are higher among the funded group. Similar results are seen for Van den Besselaar and Sandstrom (2015).

Thus, the use of multiple indicators allows sensitivity to the multidimensional aspects of research impact. While it is likely the panel of most useful indicators will vary across research programs and funding goals, the methods for integrating these variables will vary as well. Some have argued that future holistic evaluation

frameworks will need to involve qualitative and quantitative aspects of research quality and impact as well as peer and end-user evaluation to truly capture the public value of research (Donovan, 2007). In this vein, the Payback framework, which gauges "not just outputs but also outcomes derived from over a decade of investment" and takes into account the latency of impact and the attribution to multiple sources, has been suggested as best practice in research evaluation (Donovan, 2011). This framework integrates data from knowledge creation, benefits to future research, political benefits, health sector benefits and economic benefits (Bornmann, 2013). One downside to this very comprehensive approach is its labor-intensive nature and may not be relevant to assessment of individual projects. Others have focused on quantitating productive interactions between scientists and stakeholders, which is postulated to be a key generator of societal impact, although some have called for more studies to confirm this assumption (Molas-Gallart et al., 2000; Spaapen and Van Drooge, 2011; Bornmann, 2013; De Jong et al., 2014). One challenge to these types of integrations is the identification of criteria and measurable indicators for feasible assessment, and several frameworks have been suggested to address this (Sarli et al., 2010; Luke et al., 2018). Nevertheless, no standard method has been created that "can measure the benefit of research to society reliably and with validity" (Bornmann, 2017). Further, most evaluations of impact fail to take into account "inequality, random chance, anomalies, the right to make mistakes, unpredictability and a high significance of extreme events" which are hallmarks of the scientific process and likely distort any measurements of impact (Bornmann, 2017). Finally, the effect such impact assessment has on funding incentives is non-trivial, and likely influences ex-ante peer review decisions (Lindner and Nakamura, 2015; Bornmann, 2017); an important consideration when attempting to validate the peer review process.

OVERVIEW ANALYSIS OF PEER REVIEW VALIDATION STUDIES

Table 1 lists the collection of papers we identified examining the validity of peer review decisions through research outputs, which were published over the last 21 years, with a median age of 6.5 years. In general, studies had to have access to funding decisions or peer review scores or both and their relationship to external research inputs/outputs to be included. There are 48 studies included, 44% (21) are US based, 46% are European (22), 4% are Canadian (2) and 4% from Australia/New Zealand (2) and 2% from South America (1). Sample size ranged from 20 to 130,000 with a median of 828 (standard error = 3,534). 69% (33) of the studies employed bibliometrics as a research output, although several studies employed alternative measures, like project initiation and completion, patents and technology licensing, post-project peer review, levels of international collaboration, future funding success, securing tenure track positions, and career satisfaction. Collectively, 52% (25) of the studies used non-bibliometric data but only 25% (12) of all projects used more than one measure of research output. Of the studies that rely on only one indicator (36), 64% (23) rely on bibliometric measures.

Twenty-nine percent (14) are conducted without an unfunded control, and all but one of this group examines review scores and output of funded projects. Of this subset, 71% (10) provided evidence for some level of predictive validity of review decisions. Of the 29% (4) that did not, two studies used citation level per dollar spent (Danthi et al., 2014; Doyle et al., 2015) which can mask correlations, one only looked at a limited range of peer review scores, ignoring poorer scoring projects (Fang et al., 2016) and one study had a very small sample size of 40 (Scheiner and Bouchie, 2013). 71% (34) of studies listed have unfunded controls and of those, 91% (31) showed some level of predictive validity of review decisions. It has been previously suggested that another important variable in testing validity is the time window when impact is measured, especially for bibliometric impact (Van den Besselaar and Sandstrom, 2015). We find for bibliometric studies that, while most have a range, the median maximum time at which impact is measured is 5.0 $\pm$ 1.0 years after the review decision, and that 17% (3) showed no predictive validity for 5 years or less vs. 20% for more than 5 years.

It should be noted that many of the differences in impact observed were small, especially with regard to bibliometric measures. Also, several studies indicated that, when the poorer scoring unfunded applicants or poorer scoring projects were excluded from analysis, the validity disappears, although this depended on the metric used (Fang and Meyer, 2003; Hornbostel et al., 2009; Escobar-Alvarez and Myers, 2013). Also, several have noted the large degree of variability in bibliometric measures, especially with regard to projects, which obfuscate strong correlations or firm conclusions. In addition, interpretation of results was sometimes made difficult due to the potential effect of the funding itself. Nevertheless, overall these results suggest at least a coarse discriminatory power, able to separate poor proposals from better ones, but not necessarily good from great. While these results should give us pause in the current era of low funding success rates, they also suggest that more needs to be done to include a variety of external impact measures for validation studies, as well as in parallel, focusing on the internal validity of review decision making procedures.

CONCLUSIONS

It is clear that despite the importance of the peer review process in determining billions of research dollars funded in the US, there are still only a handful of studies conducted with this focus (most of which were published in the last 7 years) and less than half are US based. More research needs to be done to understand the scientific validity of this process, which means improved access to pre-funding peer review data. Academics should work with funding agencies (both federal and private funders) to negotiate agreements to gain access to this data. Funding agencies should invest in these studies.

Second, it is clear that there are many ways to identify success, and the scientometrics community has warned that multiple indicators and a well-rounded approach should be used to assess the value of research (Hicks et al., 2015). Yet, the majority of these studies here use only one type of indicator, and of those, bibliometric measures are the most used. Many issues surround the use of bibliometric measures as an accurate indicator of

TABLE 1 | Summary of literature.

Paper	Funding source	N	Scoring (S) or Funding decision (FD)	Impact indicator	Unfunded control	Impact time period (years)	Summary of results
Armstrong et al., 1997	Heart and Stroke Foundation of Canada	192	FD	Number of publications and Citations ex-post	Y	3–12 years ex-post	Funded applicants cited more than unfunded
Berg, 2011	NIH/NIGMS	789	S	Citations and Publications ex-post	N	5 years ex-post	Moderate correlation between scores and citations/publications
Bloch et al., 2014	Danish Agency for Science, Technology and Innovation	3,027	S/FD	Securing faculty positions ex post	Y	3 years ex post	Funded applicants acquire more funding and faculty positions compared to similarly scoring unfunded applicants
Bol et al., 2018	Innovation Research Incentives Scheme (Netherlands)	1255	S/FD	Ex post grant funding	Y	7–12 years ex post	Funded applicants acquire more funding over time compared to similarly scoring unfunded applicants
Bornmann and Daniel, 2006	Boehringer Ingelheim Fonds Post-doctoral Fellowship	397	FD	Citations ex-ante	Y	1–9 years ex-ante	Funded applicants have more citations than rejected applicants ex-ante
Bornmann et al., 2008b	European Molecular Biology Organization Post-doc and young investigator funds	965	FD	Citations (ex-ante and ex-post)	Y	5 years ex ante and 8 years ex post	Funded applicants higher citations than unfunded both ex-ante and ex-post (although there is significant type I/II error)
Bornmann et al., 2010	European Molecular Biology Organization	668	S/FD	Normalized citations, publications ex ante and ex post	Y	3 years ex ante and 3 years ex post	Awarded applicants more cited than rejected applicants, but best unfunded perform better as well as funded
Bromham et al., 2016	Australian Research Council Discovery Programme	18,476	S/FD	Interdisciplinarity and collaboration ex ante	Y	*In situ* measurement of application	Poorer Scores are associated with higher interdisciplinarity, even when factoring in level of collaboration is taken into account
Cabezas-Clavijo et al., 2013	Spanish National R&D Plan	2,333	S/FD	Citations and number of Publications ex-ante	Y	5 years ex ante	Accepted proposals better ex-ante PI performance than rejected PIs, but low level of correlation between scores and ex-ante productivity
Campbell et al., 2010	National Cancer Institute of Canada	685	S/FD	Citation, Publications ex post	Y	3 years ex post	Higher citations for funded vs. unfunded
Chai and Shih, 2016	Danish National Advanced Technology Foundation	4,224	FD	Patent and publication productivity ex post	Y	5 years ex post	Publication and patent productivity higher for funded for younger firms and larger projects
Claveria et al., 2000	Spanish Health Research Fund	2,744	S	Ex-post peer review of research outcomes	N	4–12 years ex-post	Ex-ante review scores significant predictor of ex-post review scores
Danthi et al., 2014	NIH/NHLBI	1,492	S	Normalized Citation ex-post	N	2 years ex-post	No Association between scores and citation impact/$

(Continued)

TABLE 1 | Continued

Paper	Funding source	N	Scoring (S) or Funding decision (FD)	Impact indicator	Unfunded control	Impact time period (years)	Summary of results
Decullier et al., 2014	French Ministry of Health	481	FD	Clinical project initiation, completion, publication ex post	Y	8–10 years ex post	Funded projects were more likely to be initiated but once initiated, completion and publication were unaffected by funding status
Doyle et al., 2015	NIH/NIMH	1,755	S	Normalized Citation ex-post	N	6–15 years ex-post	No Association between scores and citation impact/$
El-Sawi et al., 2009	Association of American Medical College (Medical Education Research)	20	FD	Survey on number of research products and level of collaboration ex post	Y	3–8 years ex post (products and collaborations)	Funded applicants had more research products and higher levels of collaboration than unfunded applicants
Escobar-Alvarez and Myers, 2013	Doris Duke Charitable Trust	1,441	S/FD	NIH funding success ex-post	Y	1–13 years ex-post	Funded applicants higher ex post NIH funding success than unfunded yet highly ranked applicants (ex-ante)
Fang and Meyer, 2003	Howard Hughes Medical Institute	867	FD	NIH funding success ex-post	Y	5–13 years ex-post	Funded applicants had higher NIH funding success than unfunded (similar ex-ante qualifications)
Fang et al., 2016	NIH	102,740	S	Citations and patents ex post	N	5 years ex post	Re-analysis of Li and Agha, 2015, amongst the higher scoring, poor correlation between scores and productivity
Galbraith et al., 2006	US Department of Defense	69	S	Success of early stage technologies (CRADA, license or SBIR funding) ex-post	N	2–3 years ex post	Reviewer scores only weakly predictive of commercial success of funded projects
Gallo et al., 2014	Anonymous (US)	227	S	Normalized Citation ex-post	N	8–15 years ex-post	Moderate correlation between scores and citation impact
Gush et al., 2017	New Zealand Marsden Fund	1,263	S/FD	Normalized citations ex-post	Y	5 years ex-post	Funding success associated with research output increases, but no correlation with review scores (all high performers)
Heggeness et al., 2018	NIH	14,276	S/FD	Ex post NIH grant funding	Y	7–19 years ex post	Funded applicants acquire more funding over time compared to similarly scoring unfunded applicants
Hornbostel et al., 2009	German Research Foundation	695	FD	Faculty positions, Career Satisfaction, Citation and publication levels ex post	Y	4 years ex ante and ex post (publications), 3 years ex post (Career)	Career satisfaction and faculty positions gained are higher among funded, but only marginally better citation levels than unfunded
Jacob and Lefgren, 2011a[Postdoc]	NIH Fellowships	12,189	S/FD	Number of Publications ex-post and active research career	Y	5 years ex post (publications and career)	Funded have slightly elevated publication rate over highly ranked but unfunded ex-post, also more active research careers, some correlation with review score and productivity

(Continued)

TABLE 1 | Continued

Paper	Funding source	N	Scoring (S) or Funding decision (FD)	Impact indicator	Unfunded control	Impact time period (years)	Summary of results
Jacob and Lefgren, 2011b [Grant]	NIH R01 Projects	54,741	S/FD	Number of Publications ex-post	Y	5 years ex post	Funded applicants have slight increase in productivity ex-post compared to highly ranked but unfunded, some correlation with review score and productivity
Langfeldt et al., 2012	Research Council of Norway	6,064	FD	Normalized citations ex-ante and ex-post, Survey, international co-authorship	Y	3–5 years ex post and 4–6 years ex ante (publications and co-authorship), 5–7 years (career)	Successful applicants more cited than unsuccessful ex-ante and ex-post, successful have higher degree of international co-authorship than unsuccessful
Lauer et al., 2015	NIH/NHLBI	6,873	S	Top 10% Normalized Citation ex-post	N	4–35 years ex-post	Modest relationship between scores and citation impact (no relationship with citation impact/$)
Li and Agha, 2015	NIH	130,000	S	Citations and Patents ex-post	N	5 years ex-post	Moderate correlation between scores and citations and patents
Mahoney et al., 2007	American Academy of Family Physicians Foundation	95	FD	Number of publications ex-post and ex ante	Y	5 years ex post and ex ante	Funded higher publication rate compared to unfunded for both ex post/ex ante
Mason et al., 2013	NIH	2,893	S/FD	NIH funding success ex-post	Y	3–31 years ex-post	Funded applicants had higher NIH funding success than highly ranked but unfunded applicants
Mavis and Katz, 2003	March of Dimes Birth Defects Foundation	439	FD	Number of Publications, Citations and additional funding success ex-post	Y	10 years ex-post	Funded applicants published more and received more citations than unfunded, as well as garnered more additional funding
Melin and Danell, 2006	Swedish Foundation for Strategic Research	40	FD	Number of publications and impact factor (ex-ante and ex-post), patents, international collaboration, acquiring additional funding	Y	3 years ex ante and 3 years ex post (5 years ex post for patents and collaborations)	Similar number of publications ex-ante and ex-post for funded/unfunded, but funded have higher average impact factor ex-post, more international collaborations ex-ante and ex-post and more success in future funding, as well as more patents
Mutz et al., 2015	Austrian Science Fund	1,689	S	Ex-post peer review of research outcomes	N (Data Imputation)	5–15 years ex post	Moderate correlation between ex-ante peer review evaluations and ex-post reviews of performance (may not represent independent observation)
Neufeld et al., 2013	European Research Council	758	FD	Ex-ante Normalized Citations	Y	6 years ex ante	Funded and unfunded applicants have similar pre-application productivity and citation impact (all high performers)

(Continued)

TABLE 1 | Continued

Paper	Funding source	N	Scoring (S) or Funding decision (FD)	Impact indicator	Unfunded control	Impact time period (years)	Summary of results
Pion and Cordray, 2008	Burroughs Wellcome Fund	619	S/FD	Securing faculty positions, R01 grant success, and publishing in top-ranked journals ex-post	Y	4–7 years (career/grant), 1–5 years for publications ex-post	Funded outperformed unfunded applicants in faculty position, R01 grant success, and publications, but differences diminished when controlled for review score
Pion and Ionescu-Pioggia, 2003	Burroughs Wellcome Fund	101	FD	Securing tenure track positions ex-post	N	1–4 years ex-post	Funded applicants secured more tenure track positions than unfunded
Reinhart (2009)	Swiss National Science Foundation	63	FD	Citation levels ex post	Y	7 years ex post	Funded applicants more cited than unfunded
Robitaille et al., 2015	European Research Council	5,100	S/FD	Number of publications, normalized citations ex-ante and ex post, interdisciplinarity	Y	27–31 years ex ante, 2–6 years ex post (publications and collaborations)	Funded had higher citations than unfunded overall and highly ranked but unfunded. Funding did not affect interdisciplinarity of output.
Sandstrom and Hallsten, 2008	Swedish Research Council	280	S	Publications and Citations ex ante	N	6 years ex ante	Review scores correlated with publication productivity
Saygitov, 2014	Russian Foundation for Basic Research	190	FD	Citations and publications ex ante and ex post	Y	5 years ex ante and 5 years ex post	Funded and unfunded applicants both have similar productivity
Scheiner and Bouchie, 2013	NSF	48	S	Citation levels ex-post	N	11 years ex-post	No correlation between scores and citation levels
Tesauro et al., 2013	NIH/NCI	184	S/FD	Ex post grant funding	Y	2–10 years ex post	Funded more likely to acquire ex post funding than similarly scoring unfunded applicants
Ubfal and Maffioli, 2011	Fund for the Scientific and Technological Research (Argentina)	496	FD	Collaboration ex post	Y	12–13 years ex post	Funded applicants had more collaboration ex post compared to unfunded
van den Besselaar and Leydesdorff, 2009	Netherlands Research Council (Social Sciences)	1,178	S/FD	Number of Publications and Citations ex-ante	Y	4–6 years ex ante	Funded more cited than unfunded; however top tier unfunded cited more than funded
Van den Besselaar and Sandstrom, 2015	Netherlands Social Science Council	260	S/FD	Citations/Top 10% publications ex-post, securing faculty positions	Y	8–10 years ex-post	Funded applicants higher performance than overall unfunded, but no difference with best performing unfunded. Funded applicants secured position more than unfunded
van Leeuwen and Moed, 2012	Netherlands Organization for Scientific Research	3,660	FD	Normalized Citations ex ante	Y	4–8 years ex ante	Funded applicants higher citations than unfunded
Wenneras and Wold, 1997	Swedish Medical Research Council	114	S	Publications and Citations ex ante	N	Unclear (probably 5 year or less ex ante)	Applicant competence scores correlated with publication impact

impact, as they can depend on many other factors unrelated to research quality (Sarli et al., 2010). More work into indicators that take into account social impact and non-bibliometric methods are also needed (Bornmann, 2013). For instance, as some have pointed out that traditional citation analysis may underestimate the true impact of clinical research (Van Eck et al., 2013); prioritizing citation counts from clinical trials or clinical guidelines may be one way to highlight translational impact (Thelwall and Maflahi, 2016). Similarly, while methodological innovations are usually well cited, getting some sense of rate of usage in a field (e.g., through the use of a survey) may give a more appropriate estimation of impact beyond what is published (Brueton et al., 2014). And as the importance of reproducibility in science cannot be overstated (Ioannidis, 2005), assessments of reproducibility (e.g., the r-factor) are currently in development (Chawla, 2018). As impact indicators are generated and validated, they should be used in review validation studies.

Third, these future studies should use a combination of metrics in order to produce a more comprehensive analysis, context and validity. Only 25% of these studies used more than one impact indicator. However, some that did found peer review decisions to be predictive of success by one measure, but much less predictive by another (Melin and Danell, 2006; Hornbostel et al., 2009). Studies show huge variability in bibliometric indicators, so they need to be supplemented to give robustness to the test for validity (Danthi et al., 2014; Gallo et al., 2014). Also, different research programs have different goals which may include both bibliometric and non-bibliometric outcomes, both should be observed to give context. Similarly, program specific context should be considered. For example, research programs can evolve over time in terms of quality of applications received and funding success rates (Gallo et al., 2014). Also, one must also consider how scientific excellence is defined and measured and how the incentivization through metrics can influence research output and the review itself (Lindner and Nakamura, 2015; Bornmann, 2017; Moore et al., 2017; Ferretti et al., 2018). Subjective definitions of excellence may not always equate to high innovation or impact, and thus the context of how the review was conducted and how reviewers were instructed to interpret excellence should be considered (Luukkonen, 2012). Once a panel of indicators is decided upon, the results should be integrated and interpreted in the context of the area of science, the goals of the research program, and the implementation of the peer review. In addition, the overall societal impact needs to be considered, as well as the inherent volatility of the scientific discovery process.

Fourth, the structure of the tests of validity vary considerably across studies, some of which lack crucial controls. For instance, examining ex-post applicant performance without comparing ex-ante performance may fail to remove the effect of funding itself. Also, for studies looking at ex ante performance as a predictor of future performance, they should take into account the Matthew effect in their interpretation, as some results show that funding less-awarded groups may actually have higher impact than more distinguished groups (Langfeldt et al., 2015; Mongeon et al., 2016), and thus reviewers choosing high ex ante performers may not always pay off. For studies examining scores vs. applicant or project output, they are usually missing crucial information about the unfunded group, which limits the ability to test validity (Lindner and Nakamura, 2015). In addition, many studies have indicated low inter-rater reliability amongst panelists (Cole and Simon, 1981) and some studies indicate that review scores and rankings are much more dependent on the individual reviewer than on the proposal (Jayasinghe et al., 2003; Pier et al., 2018). Thus, there is a need to look at the internal validity of the review process with examinations of potential reviewer bias, review structures and baselines of decision making (Magua et al., 2017). These types of internal tests of review process validity are not included in this manuscript, but are crucial for assessing other expectations of peer review (Wood and Wessely, 2003), like fairness (Lee et al., 2013), efficiency (Carpenter et al., 2015) and rationality (Gallo et al., 2016).

Finally, from the results summarized in this review, it seems that peer review likely does have some coarse discrimination in determining the level and quality of output from research funding, suggesting the system does have some level of validity, although admittedly the span of funding agencies and mechanisms included in this review complicates generalization somewhat. While it may be able to separate good and flawed proposals, discrimination amongst the top tier proposals or applicants may be more difficult, which is what the system is currently charged to do given recent funding levels (Fang et al., 2016). Nevertheless, this seems to depend on the metric used, as some studies found a high degree of discrimination when tracking career success of funded and top tier unfunded applicants (Fang and Meyer, 2003; Hornbostel et al., 2009; Escobar-Alvarez and Myers, 2013), although the effects of funding itself have to be teased out (Bol et al., 2018). Also, some level of validity was found with studies involving patents, post-funding review of outcomes and levels of collaboration as well, suggesting validity across multiple outputs. Nevertheless, as the decisions become more subjective, the likelihood for bias increases, and thus much effort must be focused on ensuring the fidelity and equity of the review process. It is likely unavoidable that some meritorious research will not be funded, putting more pressure on research funding administrators to incorporate into the final funding decisions considerations of portfolio diversification, programmatic concerns, promotion of collaborations and risk considerations (Galis et al., 2012; Janssens et al., 2017; Peifer, 2017; Wahls, 2018). These considerations, as well as the creation of new funding mechanisms (e.g., funds for early career investigators; Kaiser, 2017) should complement research into peer review processes. Given that some aspects of scientific discovery may be "fundamentally unpredictable," the development of science policies that "cultivate and maintain a healthy ecosystem of scientists rather than focus on predicting individual discoveries" may be the ideal to strive for (Clauset et al., 2017).

AUTHOR CONTRIBUTIONS

SAG and SRG contributed to the conception of the review. SAG performed the statistical analysis and did the initial gathering

of the literature. SAG wrote the first draft of the manuscript. SAG and SRG wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

ACKNOWLEDGMENTS

Thanks to the American Institute of Biological Sciences (AIBS) Scientific Peer Advisory and Review Services (SPARS) staff.

REFERENCES

Adams, J. D., Black, G. C., Clemmons, J. R., and Stephan, P. E. (2005). Scientific teams and institutional collaborations: evidence from US universities, 1981–1999. *Res. Policy* 34, 259–285. doi: 10.1016/j.respol.2005.01.014

Armstrong, P. W., Caverson, M. M., Adams, L., Taylor, M., and Olley, P. M. (1997). Evaluation of the heart and stroke foundation of Canada research scholarship program: research productivity and impact. *Can. J. Cardiol.* 13, 507–516.

Azoulay, P., Stuart, T., and Wang, Y. (2013). Matthew: effect or fable? *Manage. Sci.* 60, 92–109. doi: 10.1287/mnsc.2013.1755

Ban, T. A. (2006). The role of serendipity in drug discovery. *Dial. Clin. Neurosci.* 8. 335–344.

Beaudry, C., and Kananian, R. (2013). Follow the (industry) money–The Impact of science networks and industry-to-university contracts on academic patenting in nanotechnology and biotechnology. *Indus. Innov*. 20, 241–260. doi: 10.1080/13662716.2013.791125

Berg, J. M. (2011). *Productivity Metrics and Peer Review Scores, Continued.* NIGMS Feedback Loop (blog). Available online at: https://loop.nigms.nih.gov/2011/06/productivity-metrics-and-peer-review-scores/ (Accessed July 2018).

Bloch, C., Graversen, E. K., and Pedersen, H. S. (2014). Competitive research grants and their impact on career performance. *Minerva* 52, 77–96. doi: 10.1007/s11024-014-9247-0

Bol, T., de Vaan, M., and van de Rijt, A. (2018). The Matthew effect in science funding. *Proc. Natl. Acad. Sci. U.S.A.* 115, 4887–4890. doi: 10.1073/pnas.1719557115

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *J. Assoc. Inform. Sci. Technol.* 64, 217–233. doi: 10.1002/asi.22803

Bornmann, L. (2017). Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. *High. Educ.* 73, 775–787. doi: 10.1007/s10734-016-9995-x

Bornmann, L., and Daniel, H. D. (2006). Selecting scientific excellence through committee peer review-A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics* 68, 427–440. doi: 10.1007/s11192-006-0121-1

Bornmann, L., Leydesdorff, L., and Van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: different ways of comparing rejected to awarded applications. *J. Informetr.* 4, 211–220. doi: 10.1016/j.joi.2009.10.004

Bornmann, L., Mutz, R., and Daniel, H. D. (2008c). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *J. Assoc. Inform. Sci. Technol.* 59, 830–837. doi: 10.1002/asi.20806

Bornmann, L., Mutz, R., Neuhaus, C., and Daniel, H. D. (2008a). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics Sci. Environ. Polit.* 8, 93–102. doi: 10.3354/esep00084

Bornmann, L., Wallon, G., and Ledin, A. (2008b). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two european molecular biology organization programmes. *PLoS ONE* 3:e3480. doi: 10.1371/journal.pone.0003480

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., and Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science. *Manage. Sci.* 62, 2765–2783. doi: 10.1287/mnsc.2015.2285

Boyack, K. W., Smith, C., and Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics* 114, 449–461. doi: 10.1007/s11192-017-2609-2

Bromham, L., Dinnage, R., and Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature* 534:684. doi: 10.1038/nature18315

Brueton, V. C., Vale, C. L., Choodari-Oskooei, B., Jinks, R., and Tierney, J. F. (2014). Measuring the impact of methodological research: a framework and methods to identify evidence of impact. *Trials* 15:464. doi: 10.1186/1745-6215-15-464

Cabezas-Clavijo, A., Robinson-García, N., Escabias, M., and Jiménez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: hand in hand when assessing over research proposals? *PLoS ONE* 8:e68258. doi: 10.1371/journal.pone.0068258

Callaham, M., Wears, R. L., and Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA* 287, 2847–2850. doi: 10.1001/jama.287.21.2847

Campbell, D., Picard-Aitken, M., Cote, G., Caruso, J., Valentim, R., Edmonds, S., et al. (2010). Bibliometrics as a performance measurement tool for research evaluation: the case of research funded by the National Cancer Institute of Canada. *Am. J. Eval.* 31, 66–83. doi: 10.1177/1098214009354774

Carayol, N., and Thi, T. U. N. (2005). Why do academic scientists engage in interdisciplinary research? *Res. Eval.* 14, 70–79. doi: 10.3152/147154405781776355

Carpenter, A. S., Sullivan, J. H., Deshmukh, A., Glisson, S. R., and Gallo, S. A. (2015). A retrospective analysis of the effect of discussion in teleconference and face-to-face scientific peer-review panels. *BMJ Open* 5:e009138. doi: 10.1136/bmjopen-2015-009138

Chai, S., and Shih, W. (2016). Bridging science and technology through academic–industry partnerships. *Res. Policy* 45, 148–158. doi: 10.1016/j.respol.2015.07.007

Chawla, D. S. (2018). *Online Tool Calculates Reproducibility Scores of PubMed papers.* Science. Available online at: http://www.sciencemag.org/news/2018/01/online-tool-calculates-reproducibility-scores-pubmed-papers last accessed 4/07/18 (Accessed 22 January, 2018)

Chen, C. (2016). Grand challenges in measuring and characterizing scholarly impact. *Front. Res. Metr. Anal.* 1:4. doi: 10.3389/frma.2016.00004

Clauset, A., Larremore, D. B., and Sinatra, R. (2017). Data-driven predictions in the science of science. *Science* 355, 477–480. doi: 10.1126/science.aal4217

Claveria, L. E., Guallar, E., Cami, J., Conde, J., Pastor, R., Ricoy, J. R., et al. (2000). Does peer review predict the performance of research projects in health sciences? *Scientometrics* 47, 11–23. doi: 10.1023/A:1005609624130

Cole, S., and Simon, G. A. (1981). Chance and consensus in peer review. *Science* 214, 881–886. doi: 10.1126/science.7302566

Danthi, N., Wu, C. O., Shi, P., and Lauer, M. (2014). Percentile ranking and citation impact of a large cohort of National Heart, Lung, and Blood Institute–funded cardiovascular R01 grants. *Circ. Res.* 114, 600–606. doi: 10.1161/CIRCRESAHA.114.302656

De Jong, S., Barker, K., Cox, D., Sveinsdottir, T., and Van den Besselaar, P. (2014). Understanding societal impact through productive interactions: ICT research as a case. *Res. Eval.* 23, 89–102. doi: 10.1093/reseval/rvu001

Decullier, E., Huot, L., and Chapuis, F. R. (2014). Fate of protocols submitted to a French national funding scheme: a cohort study. *PLoS ONE* 9:e99561. doi: 10.1371/journal.pone.0099561

Didegah, F., Bowman, T. D., and Holmberg, K. (2017). On the *Differences Between Citations* and *Altmetrics*: An *Investigation* of *Factors Driving* altmetrics vs. *Citations for Finnish articles*. arXiv preprint arXiv:1710.08594.

Dinsmore, A., Allen, L., and Dolby, K. (2014). Alternative perspectives on impact: the potential of ALMs and altmetrics to inform funders about research impact. *PLoS Biol.* 12:e1002003. doi: 10.1371/journal.pbio.1002003

Dolgin (2018) *PubMed Commons Closes its Doors to Comments.* Nature Feb 02 2018. Available online at: https://www.nature.com/articles/d41586-018-01591-4 (Accessed June 4, 2018).

Donovan, C. (2007). The qualitative future of research evaluation. *Sci. Public Policy* 34, 585–597 doi: 10.3152/030234207X256538

Donovan, C. (2011). State of the art in assessing research impact: introduction to a special issue. *Res. Eval.* 20, 175–179. doi: 10.3152/095820211X13118583635918

Doyle, J. M., Quinn, K., Bodenstein, Y. A., Wu, C. O., Danthi, N., and Lauer, M. S. (2015). Association of percentile ranking with citation impact and productivity in a large cohort of de novo NIMH-funded R01 grants. *Mol. Psychiatry* 20:1030. doi: 10.1038/mp.2015.71

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A. W., Cronin, E., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* 3:e3081. doi: 10.1371/journal.pone.0003081

Ebadi, A., and Schiffauerova, A. (2015). How to receive more funding for your research? Get connected to the right people!. *PLoS ONE* 10:e0133061. doi: 10.1371/journal.pone.0133061

Editorial (2018) The serendipity test. *Nature* 554:5. doi: 10.1038/d41586-018-01405-7

El-Sawi, N. I., Sharp, G. F., and Gruppen, L. D. (2009). A small grants program improves medical education research productivity. *Acad. Med.* 84, S105–S108. doi: 10.1097/ACM.0b013e3181b3707d

Escobar-Alvarez, S. N., and Myers, E. R. (2013). The Doris Duke clinical scientist development award: implications for early-career physician scientists. *Acad. Med.* 88, 1740–1746. doi: 10.1097/ACM.0b013e3182a7a38e

Fang, D., and Meyer, R. E. (2003). Effect of two Howard Hughes Medical Institute research training programs for medical students on the likelihood of pursuing research careers. *Acad. Med.* 78, 1271–1280. doi: 10.1097/00001888-200312000-00017

Fang, F. C., Bowen, A., and Casadevall, A. (2016). NIH peer review percentile scores are poorly predictive of grant productivity. *Elife* 5:e13323. doi: 10.7554/eLife.13323

Fecher, B., Friesike, S., and Hebing, M. (2015). What drives academic data sharing? *PLoS ONE* 10:e0118053. doi: 10.1371/journal.pone.0118053

Ferretti, F., Pereira, Â. G., Vértesy, D., and Hardeman, S. (2018). Research excellence indicators: time to reimagine the 'making of'? *Sci. Public Policy* 1–11. doi: 10.1093/scipol/scy007

Firestein, S. (2015). "Funding Failure," in *Failure: Why Science is so Successful* (New York, NY: Oxford University Press), 177–204.

Fortin, J. M., and Currie, D. J. (2013). Big science vs. little science: how scientific impact scales with funding. *PLoS ONE* 8:e65263. doi: 10.1371/journal.pone.0065263

Galbraith, C. S., Ehrlich, S. B., and DeNoble, A. F. (2006). Predicting technology success: identifying key predictors and assessing expert evaluation for advanced technologies. *J. Technol. Transf.* 31, 673–684. doi: 10.1007/s10961-006-0022-8

Galis, Z. S., Hoots, W. K., Kiley, J. P., and Lauer, M. S. (2012). On the value of portfolio diversity in heart, lung, and blood research. *Am. J. Respir. Crit. Care Med.* 186:575. doi: 10.1164/rccm.201208-1437ED

Gallo, S. A., Carpenter, A. S., Irwin, D., McPartland, C. D., Travis, J., Reynders, S., et al. (2014). The validation of peer review through research impact measures and the implications for funding strategies. *PLoS ONE* 9:e106474. doi: 10.1371/journal.pone.0106474

Gallo, S. A., Sullivan, J. H., and Glisson, S. R. (2016). The influence of peer reviewer expertise on the evaluation of research funding applications. *PLoS ONE* 11:e0165147. doi: 10.1371/journal.pone.0165147

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., et al. (2011). Race, ethnicity, and NIH research awards. *Science* 333, 1015–1019. doi: 10.1126/science.1196783

Gok, A., Rigby, J., and Shapira, P. (2016). The impact of research funding on scientific outputs: evidence from six smaller European countries. *J. Assoc. Inform. Sci. Technol.* 67, 715–730. doi: 10.1002/asi.23406

Guerzoni, M., Aldridge, T. T., Audretsch, D. B., and Desai, S. (2014). A new industry creation and originality: insight from the funding sources of university patents. *Res. Policy* 43, 1697–1706. doi: 10.1016/j.respol.2014.07.009

Gurney, T., Horlings, E., Van den Besselaar, P., Sumikura, K., Schoen, A., Laurens, P., et al. (2014). Analysing knowledge capture mechanisms: methods and a stylised bioventure case. *J. Informetr.* 8, 259–272. doi: 10.1016/j.joi.2013.12.007

Gush, J., Jaffe, A., Larsen, V., and Laws, A. (2017). The effect of public funding on research output: the New Zealand Marsden Fund. *NZ Econ. Papers* 1–22. doi: 10.1080/00779954.2017.1325921

Hagen, N. T. (2008). Harmonic allocation of authorship credit: source-level correction of bibliometric bias assures accurate publication and citation analysis. *PLoS ONE* 3:e4021. doi: 10.1371/journal.pone.0004021

Heggeness, M. L., Ginther, D. K., Larenas, M. I., and Carter-Johnson, F. D. (2018). *The Impact of Postdoctoral Fellowships on a Future Independent Career in Federally Funded Biomedical Research (No. w24508).* National Bureau of Economic Research.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., and Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature* 520:429. doi: 10.1038/520429a

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102:16569. doi: 10.1073/pnas.0507655102

Hornbostel, S., Bohmer, S., Klingsporn, B., Neufeld, J., and von Ins, M. (2009). Funding of young scientist and scientific excellence. *Scientometrics* 79, 171–190. doi: 10.1007/s11192-009-0411-5

Huang, Z., Chen, H., Li, X., and Roco, M. C. (2006). Connecting NSF funding to patent innovation in nanotechnology (2001–2004). *J. Nanopart. Res.* 8, 859–879. doi: 10.1007/s11051-006-9147-9

Hutchins, B. I., Yuan, X., Anderson, J. M., and Santangelo, G. M. (2016). Relative Citation Ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLoS Biol.* 14:e1002541. doi: 10.1371/journal.pbio.1002541

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PLoS ONE* 3:e2778. doi: 10.1371/journal.pone.0002778

Jacob, B. A., and Lefgren, L. (2011a). The impact of research grant funding on scientific productivity. *J. Public Econ.* 95, 1168–1177. doi: 10.1016/j.jpubeco.2011.05.005

Jacob, B. A., and Lefgren, L. (2011b). The impact of NIH postdoctoral training grants on scientific productivity. *Res. Policy* 40, 864–874. doi: 10.1016/j.respol.2011.04.003

Janssens, A. C. J., Miller, G. W., and Narayan, K. V. (2017). The data and analysis underlying NIH's decision to cap research support lacked rigor and transparency: a commentary. *PeerJ Preprints*.5:e3106v1.doi: 10.7287/peerj.preprints.3106v1

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *J. R. Stat. Soc. A.* 166, 279–300. doi: 10.1111/1467-985X.00278

Kaatz, A., Lee, Y. G., Potvien, A., Magua, W., Filut, A., Bhattacharya, A., et al. (2016). Analysis of NIH R01 Application Critiques, Impact and Criteria Scores: Does the Sex of the Principal Investigator Make a Difference? *Acad. Med. J. Assoc. Am. Med. Coll.* 91:1080. doi: 10.1097/ACM.0000000000001272

Kaiser, J. (2017). *Updated: NIH Abandons Controversial Plan to Cap Grants to Big Labs, Creates New Fund for Younger Scientists Science June 8 2017.* Available online at: http://www.sciencemag.org/news/2017/06/updated-nih-abandons-controversial-plan-cap-grants-big-labs-creates-new-fund-younger (Accessed July 4, 2018).

Kaltman, J. R., Evans, F. J., Danthi, N. S., Wu, C. O., DiMichele, D. M., and Lauer, M. S. (2014). Prior publication productivity, grant percentile ranking, and topic-normalized citation impact of NHLBI cardiovascular R01 grants. *Circ. Res.* 115, 617–624. doi: 10.1161/CIRCRESAHA.115.304766

Keserci, S., Livingston, E., Wan, L., Pico, A. R., and Chacko, G. (2017). Research synergy and drug development: bright stars in neighboring constellations. *Heliyon* 3:e00442 doi: 10.1016/j.heliyon.2017.e00442

Knoepfler, P. (2015). Reviewing post-publication peer review. *Trends Genet.* 31, 221–223. doi: 10.1016/j.tig.2015.03.006

Langfeldt, L., Benner, M., Sivertsen, G., Kristiansen, E. H., Aksnes, D. W., Borlaug, S. B., et al. (2015). Excellence and growth dynamics: a comparative study of the Matthew effect. *Sci. Public Policy* 42, 661–675. doi: 10.1093/scipol/scu083

Langfeldt, L., Ramberg, I., Sivertsen, G., Bloch, C., and Olsen, D. S. (2012). *Evaluation of the Norwegian Scheme for Independent Research Projects (FRIPRO)* Available online at: http://www.technopolis-group.com/wp-content/uploads/2014/04/1545_RCN_Background_Report_No07_Users_Experience.pdf (Accessed July, 2018).

Lauer, M. (2015). *Perspectives on Peer Review at the NIH*. Available online at: https://nexus.od.nih.gov/all/2015/11/12/perspectives-on-peer-review-at-the-nih/ (Accessed May 4, 2018).

Lauer, M. S., Danthi, N. S., Kaltman, J., and Wu, C. (2015). 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. *Circ. Res.* 117, 239–243. doi: 10.1161/CIRCRESAHA.115.306830

Lee, C. J. (2015). Commensuration bias in peer review. *Philos. Sci.* 82, 1272–1283. doi: 10.1086/683652

Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. (2013). Bias in peer review. *J. Assoc. Inform. Sci. Technol.* 64, 2–17. doi: 10.1002/asi.22784

Leydesdorff, L., Bornmann, L., Comins, J. A., and Milojevic, S. (2016). Citations: indicators of quality? The impact fallacy. *Front. Res. Metr. Analyt.* 1:1. doi: 10.3389/frma.2016.00001

Li, D., and Agha, L. (2015). Big names or big ideas: do peer-review panels select the best science proposals? *Science* 348, 434–438. doi: 10.1126/science.aaa0185

Li, D., Azoulay, P., and Sampat, B. N. (2017). The applied value of public investments in biomedical research. *Science* 356, 78–81. doi: 10.1126/science.aal0010

Lindner, M. D., and Nakamura, R. K. (2015). Examining the predictive validity of NIH peer review scores. *PLoS ONE* 10:e0126938. doi: 10.1371/journal.pone.0126938

Luke, D. A., Sarli, C. C., Suiter, A. M., Carothers, B. J., Combs, T. B., Allen, J. L., et al. (2018). The translational science benefits model: a new framework for assessing the health and societal benefits of clinical and translational sciences. *Clin. Transl. Sci.* 11, 77–84. doi: 10.1111/cts.12495

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: emerging ERC practices. *Res. Eval.* 21, 48–60. doi: 10.1093/reseval/rvs001

Magua, W., Zhu, X., Bhattacharya, A., Filut, A., Potvien, A., Leatherberry, R., et al. (2017). Are female applicants disadvantaged in National Institutes of Health peer review? Combining algorithmic text mining and qualitative methods to detect evaluative differences in R01 reviewers' critiques. *J. Wom. Health* 26, 560–570. doi: 10.1089/jwh.2016.6021

Mahoney, M. C., Verma, P., and Morantz, S. (2007). Research productivity among recipients of AAFP foundation grants. *Anna. Fam. Med.* 5, 143–145. doi: 10.1370/afm.628

Marsh, H. W., Jayasinghe, U. W., and Bond, N. W. (2008). Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *Am. Psychol.* 63:160. doi: 10.1037/0003-066X.63.3.160

Mason, J. L., Lei, M., Faupel-Badger, J. M., Ginsburg, E. P., Seger, Y. R., DiJoseph, L., et al. (2013). Outcome evaluation of the National Cancer Institute career development awards program. *J. Cancer Educ.* 28, 9–17. doi: 10.1007/s13187-012-0444-y

Mavis, B., and Katz, M. (2003). Evaluation of a program supporting scholarly productivity for new investigators. *Acad. Med.* 78, 757–765. doi: 10.1097/00001888-200307000-00020

Melin, G., and Danell, R. (2006). The top eight percent: development of approved and rejected applicants for a prestigious grant in Sweden. *Sci. Public Policy* 33, 702–712. doi: 10.3152/147154306781778579

Merton, R. K. (1968). The Matthew effect in science: the reward and communication systems of science are considered. *Science* 159, 56–63. doi: 10.1126/science.159.3810.56

Merton, R. K., and Barber, E. (2011). The *Travels* and *Adventures* of *Serendipity: A Study in Sociological Semantics and the Sociology of Science.* Princeton, NJ: Princeton University Press.

Milat, A. J., Bauman, A. E., and Redman, S. (2015). A narrative review of research impact assessment models and methods. *Health Res. Policy Syst.* 13:18. doi: 10.1186/s12961-015-0003-1

Mischo, W. H., Schlembach, M. C., and O'Donnell, M. N. (2014). An analysis of data management plans in University of Illinois National Science Foundation grant proposals. *J. eSci. Librarianship* 3, 3. doi: 10.7191/jeslib.2014.1060

Molas-Gallart, J., Tang, P., and Morrow, S. (2000). Assessing the non-academic impact of grant-funded socio-economic research: results from a pilot study. *Res. Eval.* 9, 171–182. doi: 10.3152/147154400781777269

Mongeon, P., Brodeur, C., Beaudry, C., and Lariviere, V. (2016). Concentration of research funding leads to decreasing marginal returns. *Res. Eval.* 25, 396–404. doi: 10.1093/reseval/rvw007

Moore, S., Neylon, C., Eve, M. P., O'Donnell, D. P., and Pattinson, D. (2017). "Excellence R Us:" university research and the fetishisation of excellence. *Palgr. Commun.* 3:16105. doi: 10.1057/palcomms.2016.105

Munafo, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., et al. (2017). A manifesto for reproducible science. *Nat. Hum. Behav.* 1:21. doi: 10.1038/s41562-016-0021

Mutz, R., Bornmann, L., and Daniel, H. D. (2015). Testing for the fairness and predictive validity of research funding decisions: a multilevel multiple imputation for missing data approach using ex-ante and ex-post peer evaluation data from the Austrian science fund. *J. Assoc. Inform. Sci. Technol.* 66, 2321–2339. doi: 10.1002/asi.23315

Neufeld, J., Huber, N., and Wegner, A. (2013). Peer review-based selection decisions in individual research funding, applicants' publication strategies and performance: the case of the ERC Starting Grants. *Res. Eval.* 22, 237–247. doi: 10.1093/reseval/rvt014

Nieminen, P., Carpenter, J., Rucker, G., and Schumacher, M. (2006). The relationship between quality of research and citation frequency. *BMC Med. Res. Methodol.* 6:42. doi: 10.1186/1471-2288-6-42

NIH (2014). *Review Criteria at a Glance.* Available online at: https://grants.nih.gov/grants/peer/Review_Criteria_at_a_Glance_MasterOA.pdf (Accessed June 4, 2018).

NIH (2016). *Overall Impact Versus Significance.* Available online at: https://grants.nih.gov/grants/peer/guidelines_general/impact_significance.pdf (Accessed June 4, 2018).

NIH (2017). *Mission and Goals*. Available online at: https://www.nih.gov/about-nih/what-we-do/mission-goals (Accessed June 4, 2018).

Payne, A. A., and Siow, A. (2003). Does federal research funding increase university research output? *Adv. Econ. Anal. Policy* 3. doi: 10.2202/1538-0637.1018

Peifer, M. (2017). The argument for diversifying the NIH grant portfolio. *Mol. Biol. Cell* 28, 2935–2940. doi: 10.1091/mbc.e17-07-0462

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., et al. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2952–2957. doi: 10.1073/pnas.1714379115

Pion, G., and Ionescu-Pioggia, M. (2003). Bridging postdoctoral training and a faculty position: initial outcomes of the Burroughs Wellcome Fund Career Awards in the Biomedical Sciences. *Acad. Med.* 78, 177–186. doi: 10.1097/00001888-200302000-00012

Pion, G. M., and Cordray, D. S. (2008). The burroughs wellcome career award in the biomedical sciences: challenges to and prospects for estimating the causal effects of career development programs. *Eval. Health Prof.* 31, 335–369. doi: 10.1177/0163278708324434

Piwowar, H. A., Day, R. S., and Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2:e308. doi: 10.1371/journal.pone.0000308

Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics* 81, 789–809. doi: 10.1007/s11192-008-2220-7

Robitaille, J. P., Macaluso, B., Pollitt, A., Gunashekar, S., and Larivière, V. (2015). *Comparative Scientometric Assessment of the Results of ERC-Funded Projects.* Bibliometric Assessment Report (D5). Available online at: https://erc.europa.eu/sites/default/files/document/file/ERC_Alternative_Metrics_report.pdf (Accessed July, 2018).

Rosenbloom, J. L., Ginther, D. K., Juhl, T., and Heppert, J. A. (2015). The effects of research & development funding on scientific productivity: academic chemistry, 1990-2009. *PLoS ONE* 10:e0138176. doi: 10.1371/journal.pone.0138176

Sandstrom, U., and Hallsten, M. (2008). Persistent nepotism in peer-review. *Scientometrics* 74, 175–189. doi: 10.1007/s11192-008-0211-3

Sanyal, P. (2003). Understanding patents: the role of R&D funding sources and the patent office. *Econ. Innov. N. Technol.* 12, 507–529. doi: 10.1080/714933760

Sarli, C. C., Dubinsky, E. K., and Holmes, K. L. (2010). Beyond citation analysis: a model for assessment of research impact. *J. Med. Libr. Assoc.* 98:17. doi: 10.3163/1536-5050.98.1.008

Saygitov, R. T. (2014). The Impact of Funding through the RF President's Grants for Young Scientists (the field–Medicine) on Research Productivity: a Quasi-Experimental Study and a Brief Systematic Review. *PLoS ONE* 9:e86969. doi: 10.1371/journal.pone.0086969

Scheiner, S. M., and Bouchie, L. M. (2013). The predictive power of NSF reviewers and panels. *Front. Ecol. Envir.* 11, 406–407. doi: 10.1890/13.WB.017

Spaapen, J., and Van Drooge, L. (2011). Introducing 'productive interactions' in social impact assessment. *Res. Eval.* 20, 211–218. doi: 10.3152/095820211X12941371876742

Stevens, G. A., and Burley, J. (1997). 3,000 raw ideas= 1 commercial success! *Res. Technol. Manag.* 40, 16–27.

Sugimoto, C. R., Work, S., Lariviere, V., and Haustein, S. (2017). Scholarly use of social media and altmetrics: a review of the literature. *J. Assoc. Inform. Sci. Technol.* 68, 2037–2062. doi: 10.1002/asi.23833

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data sharing by scientists: practices and perceptions. *PLoS ONE* 6:e21101. doi: 10.1371/journal.pone.0021101

Tesauro, G. M., Seger, Y. R., DiJoseph, L., Schnell, J. D., and Klein, W. M. (2013). Assessing the value of a Small Grants Program for behavioral research in cancer control. *Transl. Behav. Med.* 4, 79–85. doi: 10.1007/s13142-013-0236-x

Thelwall, M., and Maflahi, N. (2016). Guideline references and academic citations as evidence of the clinical value of health research. *J. Assoc. Inform. Sci. Technol.* 67, 960–966. doi: 10.1002/asi.23432

Ubfal, D., and Maffioli, A. (2011). The impact of funding on research collaboration: evidence from a developing country. *Res. Policy* 40, 1269–1279. doi: 10.1016/j.respol.2011.05.023

van den Besselaar, P., and Leydesdorff, L. (2009). Past performance, peer review and project selection: a case study in the social and behavioral sciences. *Res. Eval.* 18, 273–288. doi: 10.3152/095820209X475360

Van den Besselaar, P., and Sandstrom, U. (2015). Early career grants, performance, and careers: a study on predictive validity of grant decisions. *J. Informetr.* 9, 826–838. doi: 10.1016/j.joi.2015.07.011

Van Eck, N. J., Waltman, L., van Raan, A. F., Klautz, R. J., and Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE* 8:e62395.

van Leeuwen, T. N., and Moed, H. F. (2012). Funding decisions, peer review, and scientific excellence in physical sciences, chemistry, and geosciences. *Res. Eval.* 21, 189–198. doi: 10.1093/reseval/rvs009

Van Noorden, R. (2010). Metrics: a profusion of measures. *Nature* 465, 864–866. doi: 10.1038/465864a

Van Tuyl, S., and Whitmire, A. L. (2016). Water, water, everywhere: defining and assessing data sharing in academia. *PLoS ONE* 11:e0147942. doi: 10.1371/journal.pone.0147942

Wahls, W. P. (2018). Point of View: the NIH must reduce disparities in funding to maximize its return on investments from taxpayers. *Elife* 7:e34965. doi: 10.7554/eLife.34965

Wang, D., Song, C., and Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science* 342, 127–132. doi: 10.1126/science.1237825

Warren, H. R., Raison, N., and Dasgupta, P. (2017). The rise of altmetrics. *Jama* 317, 131–132. doi: 10.1001/jama.2016.18346

Wenneras, C., and Wold, A. (1997). Nepotism and sexism in peer-review. *Nature* 387:341. doi: 10.1038/387341a0

Wood, F., and Wessely, S. (2003). "Peer review of grant applications: a systematic review," in *Peer Review in Health Sciences*, eds Godlee and Jefferson (London: BMJ Publications), 14–31.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.