Assessing Value in Biomedical Research:

The PQRST of Appraisal and Reward

John P.A. Ioannidis, MD, DSc and

Departments of Medicine and Health Research and Policy, Stanford University School of Medicine, Palo Alto, California

Department of Statistics, Stanford University School of Humanities and Sciences, Palo Alto, California

Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Palo Alto, California

Muin J. Khoury, MD, PhD

Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia

National Cancer Institute, National Institutes of Health, Bethesda, Maryland

Production of scientific work is regulated by reward systems. Scientists are typically rewarded for publishing articles, obtaining grants, and claiming novel, significant results. However, emphasis on publication can lead to least publishable units, authorship inflation, and potentially irreproducible results. Emphasis on claiming significant results leads to lack of publication of nonsignificant high-quality studies or to massaging data to obtain "positive" results. Emphasis on novelty leaves no incentives to spend resources on replicating prior findings to probe their correctness. Data owners have a publishing advantage without incentives to share with competitor scientists.

In the past, grapevine knowledge among the few knowledgeable experts allowed discerning the good work from the waste. But currently the noise-to-signal ratio is tremendous, with the proliferation of technologies (such as genomics) and journals. Thousands of new journals publish work for a fee, regardless of the quality of the work.[1] To change the tide, the criteria by which scientists and their teams are rewarded for their efforts by agencies that fund them and institutions that host them should be revisited,[2] aligning criteria with the desired outcomes: research that is productive, high-quality, reproducible, shareable, and translatable–or PQRST for short.

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, Stanford University, Stanford, CA 94305, (jioannid@stanford.edu).

Disclaimer: The views expressed in this Viewpoint are those of the authors and do not reflect the official opinion of the Centers for Disease Control and Prevention or the National Institutes of Health.

Productivity metrics should reward high-influence science rather than least publishable units and decrease publication bias against negative results. Instead of counting each and every publishable unit, even now several major universities ask only for the top papers from each candidate for appointment or promotion. However, the process can be standardized. Citation databases such as ISI Web of Knowledge/Essential Science Indicators automatically identify scientific fields and the x% top-cited articles in each scientific field and each year (x can be set at 1%, 10%, or other desirable percentage). Authorship contributions should also be considered when allocating credit for multiauthored papers using standard formulas such as harmonic adjustments.[3]

Another useful metric is the proportion of published scientific work emanating from a research project. Funders can keep records of the publication of funded projects. Even for clinical trials funded by federal resources, a substantial proportion remain unpublished several years after their completion.[4] For other types of research, nonpublication is likely to be even more frequent. Should investigators receive more funds if they used previous funds without publishing anything? In fields such as clinical research of interventions, public study registration has become widely accepted (and even enforced by regulatory agencies) and the concept of registration can be extended to other fields when appropriate.[5] Registration allows documenting whether research studies result in published reports within a reasonable time of completion. Registration of protocols and analysis plans can also help evaluate selective reporting, ie, whether only some outcomes or analyses have been published and whether analyses deviate from promised plans.

For quality assessments, focusing on top-cited articles in assessing productivity already captures some aspects of quality, but citations and quality are not perfectly correlated. Funders can ask that protocols and studies fulfill specific quality standards in their design, implementation, and analysis. The National Institutes of Health is already moving in this direction with implementation of checklists.[6] There are also numerous reporting standards for diverse research fields, as summarized by the EQUATOR initiative.[7] Each field should agree on what quality features are essential. Nevertheless, difficulties in rating quality objectively should not be underestimated. Under pressure to comply, investigators may simply check off requested items while ignoring other fundamental issues relevant to the specific study. Quality assessments may focus on very few, uncontroversial, and easily verifiable study aspects. These assessments also may be used to promote improvement of that particular aspect in the whole field, eg, routine use of randomization and blinded assessments in preclinical animal studies.[8]

Reproducibility can range from repeating analyses with raw data to independent replication using different materials or study participants or even study designs different (more rigorous) than the original study. Some types of reproducibility checks are easy. Others are prohibitively difficult, eg, performing another similar trial with new participants and 10 years of follow-up to independently replicate the results of a clinical study. This should be taken into account in deciding whether replication and reproducibility checks should be requested routinely (eg, when easy and inexpensive to do) or under select circumstances (eg, only for the most influential papers, if difficult and expensive to perform).

Sharing can also be measured. For each scientist, it is possible to assess how many papers are accompanied by shareable data, materials, or protocols. Indexing databases such as PubMed and ClinicalTrials.gov could note in each new publication record whether such shared resources are available. Many funders and journals have already made sharing practices mandatory for particular research types.

As for translation, much excellent research has no recognizable translational application or benefit. Scientific influence, if any, often becomes manifest many years after the initial discovery. However, translational performance is relevant in research with direct applied aspirations, and this covers most of preclinical investigation and clinical medicine. For example, for preclinical studies of interventions, the translational milestone may be successful evaluation of the same intervention in humans; for clinical research, it may be licensing or approval for clinical use.

Given current resources, some of these indices can be easily evaluated for all scientists and for all their work. Other metrics and investigators need more focused appraisals; eg, it is impossible to perform reproducibility checks on every single published article. It is much easier to focus on the most influential articles, which are the ones considered in item P (eg, the top-cited 1%). A small budget is needed to reproduce articles that attain top scientific influence and thus have a major effect on the course of science. Many such articles already include replications, eg, currently all most-cited genome association studies include by default extensive replication in independent populations. Assessments of quality and translational influence that lack all-encompassing automated databases may also need to focus on the most influential work. Proportion of published work and assessments of sharing practices can relatively easily be automated science-wide by funders, registries, indexing online libraries, or other resources.

The Table illustrates some suggestions for how to potentially operationalize these principles. The suggestions are not prescriptive but may offer ideas to funders and other stakeholders for next practical steps. The exact combination or weighting of indices will require discussion and consensus among stakeholders. Selected reward system choices should also factor the potential for gaming any appraisal and reward system. Potential untoward consequences of gaming should be anticipated, minimized, and monitored. For example, scientists may acknowledge funding from specific grants for entirely unrelated published work if their career depends on demonstrating that funding resulted in publications. If so, funders should verify that the funded work has indeed been published. Or, if reward is given only for top-cited articles, networking between investigators and journals may create a citation factory of mediocre articles that mutually propel themselves toward the top-cited range. Some of the other indices will correct this; eg, observational nutritional epidemiology has some of the most-cited papers across all science, but much of this work has failed replication.

In addition, funding agencies, universities, research institutions, academies, professional societies, and prestigious award organizations may also have PQRST indices based on the research work they sponsor or perform and the scientists behind this work. Further discussion is needed among stakeholders to refine these indices and to evaluate them within

each scientific field. Special consideration will also be needed in rewarding research based on transdisciplinary team science that crosses the boundaries of multiple scientific disciplines.

Acknowledgments

References

1. Bohannon J. Who's afraid of peer review? Science. 2013; 342(6154):60–65. [PubMed: 24092725]

2. Macleod MR, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. Lancet. 2014; 383(9912):101–104. [PubMed: 24411643]

3. Aziz NA, Rozing MP. Profit (p)-index: the degree to which authors profit from co-authors. PLoS One. 2013; 8(4):e59814. [PubMed: 23573211]

4. Gordon D, Taddei-Peters W, Mascette A, Antman M, Kaufmann PG, Lauer MS. Publication of trials funded by the National Heart, Lung, and Blood Institute. N Engl J Med. 2013; 369(20):1926–1934. [PubMed: 24224625]

5. Dal-Ré R, Ioannidis JP, Bracken MB, et al. Making prospective registration of observational research a reality. Sci Transl Med. 2014; 6(224):224cm1.

6. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014; 505(7485):612–613. [PubMed: 24482835]

7. Simera I, Moher D, Hoey J, Schulz KF, Altman DG. A catalogue of reporting guidelines for health research. Eur J Clin Invest. 2010; 40(1):35–53. [PubMed: 20055895]

8. Macleod M. Why animal research needs to improve. Nature. 2011; 477(7366):511. [PubMed: 21956292]

Table

PQRST Index for Appraising and Rewarding Research.

Item in PQRST Index	Operationalization	
	Example	Data Source
P (productivity)	Number of publications in the top tier % of citations for the scientific field and year	ISI Essential Science Indicators (automated)
	Proportion of funded proposals that have resulted in 1 published reports of the main results	Funding agency records and automated recording of acknowledged grants (eg, PubMed)
	Proportion of registered protocols that have been published 2 y after the completion of the studies	Study registries such as ClinicalTrials.gov for trials
Q (quality of scientific work)	Proportion of publications that fulfill 1 quality standards	Need to select standards (different per field/design) and may then automate to some extent; may limit to top-cited articles, if cumbersome
R (reproducibility of scientific work)	Proportion of publications that are reproducible	No wide-coverage automated database currently, but may be easy to build, especially if limited to the top-cited pivotal papers in each field
S (sharing of data and other resources)	Proportion of publications that share their data, materials, and/or protocols (whichever items are relevant)	No wide-coverage automated database currently, but may be easy to build, eg, embed in PubMed at the time of creation of PubMed record and update if more is shared later
T (translational influence of research)	Proportion of publications that have resulted in successful accomplishment of a distal translational milestone, eg, getting promising results in human trials for intervention tested in animals or cell cultures, or licensing of intervention for clinical trials	No wide-coverage automated database currently, would need to be curated by appraiser (eg, funding agency) and may need to be limited to top-cited papers, if cumbersome