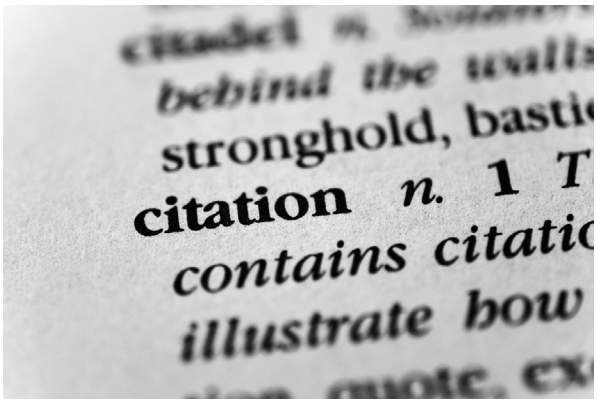
Citation metrics for appraising scientists: misuse, gaming and proper use

We need informative citation metrics that will be less prone to misuse and gaming

Citation and other metrics are widely misused, but when properly used, they can be valuable. Science itself thrives on quantitative measurement. Quantitative indicators aim to provide objective data instead of biased beliefs. Here, we focus on citation metrics in appraising scientists[1] for hiring, promotion, tenure, funding, selection for some award, recognition or bonus, or other reasons. Many tricks exist to game citation metrics (Box); however, proper use of metrics may overcome these deficiencies. Generic challenges that we describe here may partly apply also to larger, more composite entities such as the appraisal of journals, institutions or large research portfolios, for example, at a national level.

Let us assume that a gold standard could theoretically exist so that a scientist could be accurately ranked in the X percentile of impact among all fellow scientists. Then, the ratio of the number of scientists who are better than this scientist versus those who are worse than this scientist would be $R = X \div (100 - X)$. Of course, gold standards are utopian. Moreover, typically, we are not interested to rank a scientist abstractly against all scientists in the world, but against a relevant comparator group where a comparison is fair and something specific is at stake. The question is: how much off might we be in our assessment? We argue that, realistically, we should usually be happy if a metric comes within fourfold of the true R. Then, a top 1% scientist may be ranked as being anywhere in the top 4% range. An average scientist (50th percentile) may find themselves ranked between the 20th and 80th percentiles. A scientist in the lower 20% may be ranked somewhat lower or higher but will not find themselves appraised as belonging to the upper half.

In most circumstances, citation metrics are more noise-prone than this, especially if they are calculated with poorly collected or limited information. For instance, if a citation database does not cover well a scientific field, citation analyses using that database become spurious and comparing scientists with counts derived from different databases is erroneous.[2] For young scientists who have just published their first few articles, ranking based on impact metrics is not informative. Conversely, for scientists who have sufficiently lengthy careers behind them, realistically attainable precision may suffice for guiding decisions where differences between compared scientists are substantial, although it may still be insufficient for other important decisions. For example, no metrics can perfectly identify Nobel laureates. Using the best metrics, most (not all) of the Nobel laureates are in the top 0.1% of scientists,[3] but this 0.1% group still corresponds to 35 000 top scientists among the 35 million who author articles indexed in Scopus.



Metrics do not exist in a vacuum. They are typically used by experts who also try to judge the work of evaluated scientists, hopefully in sufficient informational depth, regarding the subject matter, the scientific quality and the methodological rigour of their work. Here the tension arises as experts are piled against metrics, while their relationship should be symbiotic.

Assessors who produce first rate work themselves may choose other people with first rate achievements, while evaluators with second rate achievements tend to choose people with third rate work.[4] Experts in various evaluations are often second, third or unfathomable bottom rate in their own achievements. Many environments and many institutions are non-meritocratic. This is not surprising: thinking of the normal distribution, truly excellent people and institutions are a minority, the bulk are in the middle and there are several who are really bad but, nevertheless, may still have much power and expert nay-say in their environment. Metrics can and should be used to counterbalance the situation where mediocre evaluators would otherwise be unaccountable. At the other end of the spectrum, even when the very best experts in the world are involved in decisions, absolutely perfect ranking still remains utopian. For example, hundreds and possibly thousands of scientists may be almost equally worthy of the Nobel Prize each year, but only one to three are selected. Even for more mundane, daily decisions, such as hiring in new positions, in competitive institutions dozens of excellent scientists apply, but still only one is hired. Having more information, in the form of metrics, does not hurt when one is facing close calls. Metrics do not need to antagonise programmatic needs and can actually inform them.

As metrics become widely popular and publicly visible, it is more difficult for mediocre stakeholders who do not like them to suppress them entirely. Instead, evaluators and institutions of second rate quality use

John PA Ioannidis[1]

Kevin W Boyack[2]

1 Stanford Prevention Research Center, Stanford University, Stanford, CA, United States

2 SciTech Strategies, Albuquerque, NM, United States.

jioannid@stanford.edu

doi: 10.5694/mja2.50493

Glossary of common gaming mechanisms for citation metrics

Mechanisms	Description
Predatory and other easy journals	Probably close to 10 000 journals exist worldwide, increasing at a 5% rate annually; approximately 11 000 journals are indexed by Web of Science. Far more (unknown exact number) are indexed by Google Scholar. Many journals use predatory practices and will publish anything for a fee without any review. Many others have weak review processes
"Salami slicing"	A study or analysis is divided into multiple "least publishable units", each published as a separate article. Certainly, many secondary publications have a scientific reason of existence, but many others do not
Inflated self-citations	Self-citations to prior work are necessary to place new work in context and it is even unethical to not self-cite and, thus, make new work seem more novel than it really is. However, self-citation can get out of proportion. One needs to adjust for field (small fields have higher rates of justified self-citations), the stage of career (young researchers have higher proportions of self-citations), and adjust for productivity (prolific authors will have justifiably larger absolute numbers of self-citations)
Citation farms	A group of scientists agrees secretly that they will cite massively each other's articles without proper scientific rationale. Sometimes this results in nonsensical articles that are just laundry lists of such farm citations. Citation cartels have been described also for journals that thus try to mutually increase their impact factors. Journal-level and scientist-level cartels and farms may overlap; for example, these authors may also be members of the journals' editorial boards
Coercive self-citation	Editors and/or reviewers force authors to revise their article including citations to articles published by the journals and/or authored by the editors and/or reviewers even though the citations are not really relevant to the work
Gift (honorary) authorship	Gift authorship may occur in various settings where scientists who are powerful, senior figures force young colleagues to include them as authors in many or even all of their manuscripts. Some professor/chairperson curricula vitae in some fields show an extraordinary acceleration of productivity coinciding with the time these senior players assume the directorship of large clinics and/or institutes; that is, when they are least likely to have time to be genuine authors given the administrative burden. Another common form of co-authorship is in articles that are written by industry ghost authors (whose names do not appear) and are then having power figures placed as authors so that the work can acquire more prestige in professional circles
Inflated multi-authorship	Team work and consortia can greatly enhance scientific progress. However, some consortia have very lax rules for including authors in their articles. Different consortia may vary tenfold or more in the number of authors they use for projects that otherwise require the same amount of work. The practice of inflated multi-authorship may be combined with "salami slicing" and other forms of citation gaming within a consortium

third and fourth rate metrics and they are also more likely to manipulate indicators so as to bend them to their needs and make them look compatible with their whims. It is a vicious circle: as metrics become important, there is more pressure to game their system, according to Goodhart's law.[5] Many metrics are easy to game, and some are easier to game than others. The gaming process may harm science as it fuels resources towards satisfying the gaming needs without meaningful scientific output. The number of publications, for example, is currently extremely easy to game, as there are thousands of journals (many of them unnecessary, predatory publication venues) that can publish one's work.[6] Similarly, the sum of impact factors of one's publications can also be gamed. The impact factor of single journals is only slightly more difficult to game.[7] Even total citations can be boosted by self-citation,[8] citation farms,[9] coercive citations,[10] gift co-authorship,[11] spurious extensive multi-authorship[12] — with hundreds of authors where only one or two people really deserve authorship per Vancouver criteria — and several other tricks (Box). In the following sections, we focus on specific solutions to this gaming conundrum.

Abandon metrics that have been thoroughly gamed and carefully fix those that can still be fixed

The proliferation of sophisticated tricks to game impact metrics has an analogy to computer viruses. There are always new ones emerging. It is fascinating how much ingenuity people demonstrate to manufacture nice-looking curricula vitae and associated metrics. The result is that some metrics have become so massively infected by these viruses, that the scientific community should better abandon them entirely and let the infected carcass rot away. Traditionally defined impact factor and its sum as well as raw number of publications probably belong to this category. Impact factor was originally intended to help institutional libraries select between journals for their collections. Its use, or misuse, for evaluation of researchers is not what it was designed for. Some metrics can still be salvaged, provided that proper antivirus treatment is undertaken. For example, total citations are salvageable, if one can also capture separately self-citations; identify citation farms; adjust properly for field; use accurate, standardised data; and try to understand also what is the relative contribution of a scientist in the articles they have authored.[13]

Try to adjust for specific contributions or at least for co-authorship

Clarity in contributions to scientific work is a wonderful idea,[14] but contribution attributions vary across fields[15] and can also be gamed. Moreover, many articles do not list specific contributions of each author

anyhow. One is often left with long lists where the only thing for certain is that the implausibly multi-authored team has massively violated Vancouver criteria. New antivirus solutions can be used here, such as using metrics that adjust for co-authorship and/or incorporating metrics that use author order, as a surrogate (even if imperfect one) for weight of contribution.

Read carefully the articles and read properly the metrics using the Leiden manifesto guidance

In appraising a curriculum vitae, one needs to read carefully the articles of the evaluated scientist. Similarly, one has to read carefully the metrics measuring the impact of this scientist. The Leiden manifesto[2] — a set of ten principles that have been proposed for the evaluation of research performance with metrics (www.leidenmanifesto.org) — provides a good starting point for those who want to read metrics properly, as it highlights areas that can be common sources of misinformation and misappraisal. Metrics need to be placed in the context of the mission of the researcher and the purpose of the evaluation. They need to be accurate, verifiable, open and transparent, standardised for field, accounting for the coverage of the databases they are derived from, avoiding implausible concreteness and precision, pre-emptively considering systemic effects, remaining well updated, and considering multiple, complementary indicators rather than a single silver bullet.

Train to better understand metrics

Meeting the specifications of the Leiden manifesto is a very high order task and few people are well trained to do this properly. However, training in using metrics can be a rewarding experience. The solution to all these challenges is not to abandon all indicators. Metrics will continue to circulate widely regardless of how much we complain about them. We need better metrics, not necessarily fewer. We need to scrutinise metrics very carefully. Sometimes, their careful reading can be highly informative about the culture, values, performance and practices of a scientist, their group, or institution.

Conclusion

Given that metrics cannot just go away, lack of appreciation and lack of training in what they mean and what they can and cannot tell us may be the greatest threat associated with them currently.

Competing interests: No relevant disclosures.

Provenance: Commissioned; externally peer reviewed. ■

References are available online.

1 Cronin B, Sugimoto CR, editors. Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact. Cambridge: MIT Press, 2014.

2 Hicks D, Wouters P, Waltman L, et al. Bibliometrics: the Leiden Manifesto for research metrics. *Nature* 2015; 520: 429–431.

3 Ioannidis JP, Klavans R, Boyack KW. Multiple citation indicators and their composite across scientific disciplines. *PLoS Biol* 2016; 14: e1002501.

4 Simone JV. Understanding academic medical centers: Simone's Maxims. *Clin Cancer Res* 1999; 5: 2281–2285.

5 Biagioli M. Watch out for cheats in citation game. *Nature* 2016; 535: 201.

6 Beall J. Predatory publishers are corrupting open access. *Nature* 2012; 489: 179.

7 Ioannidis JPA, Thombs BD. A user's guide to inflated and manipulated impact factors. *Eur J Clin Invest* 2019; 49: e13151.

8 Van Noorden R, Singh Chawla D. Hundreds of extreme self-citing scientists revealed in new database. *Nature* 2019; 572: 578–579.

9 Davis P. The emergence of a citation cartel. *The Scholarly Kitchen* 2012; 10 Apr http://scholarlykitchen.sspnet.org/2012/04/10/emergence-of-a-citation-cartel (viewed Sept 2019).

10 Wilhite AW, Fong EA. Coercive citation in academic publishing. *Science* 2012; 335: 542–543.

11 Mowatt G, Shirran L, Grimshaw JM, et al. Prevalence of honorary and ghost authorship in Cochrane reviews. *JAMA* 2002 Jun 5; 287: 2769–2771.

12 Ioannidis JPA, Klavans R, Boyack KW. Thousands of scientists publish a paper every five days. *Nature* 2018; 561: 167–169.

13 Ioannidis JPA, Baaas J, Klavans R, Boyack KW. A standardized citation metrics author database annotated for scientific field. *PLoS Biol* 2019; 17: e3000384.

14 Rennie D, Yank V, Emanuel L. When authorship fails. A proposal to make contributors accountable. *JAMA* 1997; 278: 579–585.

15 Sauermann H, Haeussler C. Authorship and contribution disclosures. *Sci Adv* 2017; 3: e1700404. ■