Impact Factor and other metrics for evaluating science: essentials for public health practitioners

Angelo G. Solimini¹, Renata Solimini²
¹Department of Public Health and Infectious Diseases, Sapienza University of Rome, Italy; ²Istituto Superiore di Sanità, Rome, Italy
Correspondence to: Angelo G. Solimini, Department of Public Health and Infectious Diseases, Sapienza University of Rome, Piazzale Aldo Moro 5, 00161, Roma, Italy. Email: angelo.solimini@gmail.com

Abstract
The quality of scientific evidence is doubly tied with the quality of all research activities that generates it (including the “value” of the scientists involved) and is usually, but not always, reflected in the reporting quality of the scientific publication(s). Public health practitioners, either at research, academic or management levels, should be aware of the current metrics used to assess the quality value of journals, single publications, research projects, research scientists or entire research groups. However, this task is complicated by a vast variety of different metrics and assessment methods. Here we briefly review the most widely used metrics, highlighting the pros and cons of each of them. The rigid application of quantitative metrics to judge the quality of a journal, of a single publication or of a researcher suffers from many negative issues and is prone to many reasonable criticisms. A reasonable way forward could probably be the use of qualitative assessment founded on the indications coming from few but robust quantitative metrics.

Key words: peer review, bibliometric indicators, impact factor, h-index, eigenvalue, webometric indicators, g-index

Introduction
In many cases, public health decisions are (or, at least, should be) based on the best scientific evidence available. Often, but not always, the best scientific evidence is found in publications presented in highly ranked journals. This also reflects the adoption of quality standards in reporting scientific information by some of those journals (see for example [1-3]). The quality of scientific evidence is doubly tied with the quality of all research activities that has generated it (including the “value” of the scientists involved) and usually, but, again, not always, is reflected by the reporting quality of the scientific publication(s). For example, decisions on priorities given to funding of different public health projects are also related to the quality of the projects’ proponents. For this reason, public health practitioners, either at research, academic or management levels, should be aware of the current metrics used to assess the quality value of journals, single publications, research projects, research scientists or entire research groups.

Pushed by the generalized diminishing funds for research, science evaluation is becoming more and more important to guide decisions on resource allocation and to manage the scarce available resources transparently. The methods to evaluate the quality of science are either qualitative or quantitative [4,5].

The qualitative method includes the peer review process and the evaluation by expert panels. The peer review is notoriously applied during the submission process of a manuscript for publication in a journal, while the experts’ evaluation is often applied for grading projects submitted in response to funding calls or for selecting candidates for tenure in academic or other research positions. Being subjective methods of evaluation, qualitative methods are prone to many biases. Some journals use more rigorous quality checks than others, but overall peer-review is not perfect [6]. For example, reviewers might be more prone to reject papers not in line with current paradigms or for not citing what they (and not the author) believe to be crucial publications. Flawed papers may get through while many excellent papers that only need some improvements are rejected by highly competitive journals based on priority considerations [6]. Or a review panel might value the personal connections held by candidates in a position (or
The correct name of this indicator is *Journal Impact Factor* as Garfield and Sher aimed at setting up “a simple means of comparing the quality of small journals with large ones”. They studied “the chronological distribution of citations, especially in biochemistry and molecular biology and observed that about 25% of citations referred
to articles published just a few years back. So, they concluded that the 2-year impact factor could be a remarkable predictor of future journal performance”. Afterwards, the Journal Citation Reports (JCR) - a database of journals reporting IF values and other indices such as Immediacy Index and Cited Half Life - was developed, becoming a standard tool for library science and publishing activities purposes [21].

The IF formula of a Journal X in year 2010 is calculated as follows: \( \text{IF}(X, 2010) = \frac{B}{C} \) where:

- \( A \) = sum of total citations of items published in 2010 by Journal X present in any items of any journals in Thomson database (e.g. indexed journals)
- \( B \) = 2010 citations of Journal X items published in 2008-09 (this is a subset of A)
- \( C \) = number of primary research and review articles published in 2008-09 into Journal X

Derived from IF, are 2 other bibliometric indicators: the Immediacy Index and the Cited Half Life index. The immediacy index is calculated based on the papers published in a journal in a single calendar year.

For example, the 2010 Immediacy Index of Journal X would be calculated as follows:

\[ \text{Immediacy Index (X) 2010} = \frac{A}{B} \]

- \( A \) = the number of times items published in 2010 into Journal X were cited in indexed journals during 2010
- \( B \) = the number of articles, reviews, proceedings or notes published in 2010 by Journal X.

The Journal Cited half-life is the median age of articles that were cited in a given year. Cited half-life shows how quickly articles published in a given journal, taken as a set, cease to be cited.

The IF overcomes the shortcomings of absolute (or total) citation counts. It eliminates some of the biases of such counts which favor large journals over small ones, or frequently issued journals over less frequently issued ones, and of older journals over newer ones. Many biases of IF have been suggested in the past few years ([22-98]

Textbox 1. The most frequently reported criticisms of the IF [23-25].

- IF does not capture the multidimensional phenomena of a journal’s influence on science;
- Confusion over the definition of citable items (the denominator in the calculation; see also [24]) and a lack of transparency by Thomson on this issue [23];
- IF inflated by citations to article types (such as editorials or letters), not accounted for in the denominator [24];
- Review journals often have higher impact factors and thus have an advantage over non-review journals;
- Cross field comparisons meaningless because of differences in absolute IF from one discipline to another;
- Subjective definition of disciplines included into JRC fields
- The two-year citation window is too short and penalizes some fields and their journals
- Multidisciplinary journals, which offer a mixed set of papers in terms of fields, produce a “mixed” IF of little use;
- No IF available if the journal is not indexed by Thomson Reuters;
- Journal coverage is biased against certain nations and languages (like English)
- IF is biased by citation in negative or about retracted articles [23]
- IF is biased by the “Blockbuster paper” effect ([23]; see text)
- There is a skewed distribution of citations in most fields
A concise list of IF reported biases is shown in textbox 1. However, it is worth mentioning that comparisons of JIF between disciplines should take into account the fact that journal diffusion among readers (e.g. potential citers) greatly differs among subject categories or journal types (generic journals will physiologically have more readers than specialist ones). An effort to normalize impact factors for different fields has been carried out [26,27]. What emerged was the “Median Impact Factor” (MIF), i.e. the median value of all journal impact factors in the subject category [27], and it is also included in the JCR.

Although designed for evaluating the impact of journals within a discipline, the IF is used to evaluate the impact of single articles, scientists, research groups and even entire departments. However, even if a scientist has his/her own article published in a high IF journal, this does not necessarily mean that it will have an equally high impact on the scientific community. Consider the case of an author that publishes an article in a highly impacted journal but his article totalizes zero citations in the successive years. This article has zero impact on science (apart from exceptions derived from articles cited long time after being published [28]), however, the article’s author profits from the citation (and IF) tracking effect of “blockbuster” papers [23], the ones having the “real” impact on the scientific community.

Often, in highly ranked journals, a quarter of all IF citations can be attributed to those few “star articles” published each year [29].

**The h-Index**

As already mentioned, the Impact Factor is not representative of the impact of the individual journal articles or of the individual researcher impact. In 2005, Jorge Hirsch proposed the h-index, a number that combines publication activity and citation influence, attempting to measure the productivity and impact of the work published by a scientist. It is defined as follows: “A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np − h) papers have at most h citations each.” For example, a researcher with 9 publications that each have 9 or more citations has an h-index of 9 [30]. The h-index can be calculated for free by several tools like the scHolar Index Index (http://interaction.lille.inria.fr/~roussel/projects/scholarindex/index.cgi); QuadSearch (http://quadsearch.csd.auth.gr/index.php?s=2&lan=1); Publish or Perish (http://www.harzing.com/pop.htm), that also provides g-index, citations per paper and other indices, obtaining citations from Google Scholar, or with fee subscription by the Web of Knowledge or Scopus.

Notably, the h-index of a researcher will differ based on the database used as indicated in Figure 1 for one of the authors of this paper at the time of writing.

![Figure 1](image-url)

**Figure 1.** Example calculations of the h-index and effect of the database used as source of citations for one of the authors of this paper at the time of writing.

Open and closed quadrates are citation counts using Scopus and Google Scholar respectively. Scopus retrieved 38 cited documents, Google Scholar 46. The h-index (Scopus) = 9; the h-index (Google Scholar) = 10.
of writing. As shown in figure 1, the use of Google Scholar as a citation source database will give a higher h-index than Scopus. In fact the h-index in Scopus = 9 while the h-index in Google Scholar = 10. Moreover, 8 publications were present in Google Scholar but not present in Scopus and 4 were present in Scopus but not in Google Scholar.

Reported biases and disadvantages of Google Scholar are its inclusion of non-scholarly citations, double counting of citations, less frequent updating, uneven coverage across disciplines and less comprehensive coverage of older publications/citations [31]. Researchers that are also working in highly visible organizations or research groups will have a higher probability of being noticed and their work cited over researchers that produce similar quality papers but are less in the science spotlight. Another bias reflects the already mentioned citing misbehaviour when an author prefers to cite a given researcher or article because his or her work represents the current “gold standard” in that authors' discipline ignoring the less dogmatic work of others so as not to displease the reviewers.

Other biases refer to the number of authors per paper (the h-index tends to favor disciplines with larger groups); the different citation practices in different fields (the numbers of citations per paper are different among disciplines); the dependency on the stage of a researcher’s career as h-index penalizes young scientists, i.e. the short careers, being bounded by the total number of publications.

The h-index has also been recently proposed for journals [31]. Analogous to its use for authors, the h-index for journals may provide a robust measure of sustained and durable performance of journals, rather than articles.

The g-index

The g-index was introduced by Leo Egghe in 2006. The g-index is a modification of the h-index that takes into account the presence of highly cited papers beyond the h value. It is defined as follows: “The g-index is introduced to measure the global citation performance of a set of articles. If this set is ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received (together) at least g² citations. It is shown that the g-index inherits all the good properties of the h-index and, additionally, takes better into account the citation scores of the top articles. This yields a better distinction between, and order of, the scientists from the point of view of visibility” [32].

Both the h-index and the g-index should complement each other to represent a global value of a scientist.

Eigenfactor

The Eigenfactor score measures the total influence of a journal using network based ranking systems. It was ideated by the University of Washington, on the view that the scholarly literature forms a vast network of academic papers connected to one another by citations in bibliographies and footnotes [33]. It is published in the JCR together with the other impact indices reported above. Notably, eigenfactor analysis aims at answering the question “what should I read to pursue my research?” by finding items that are relevant (a matching problem) and of sufficient high quality (a ranking problem) [34].

The algorithm implemented by eigenfactor is similar to Google’s PageRank algorithm. Journals are rated according to the number of incoming citations, with citations from highly-ranked journals weighted to make a larger contribution to the eigenfactor than those from poorly-ranked journals [35]. As a measure of importance, the Eigenfactor score scales with the size of a journal. All else equal, larger journals have larger Eigenfactor scores. As such, Eigenfactor scores are not directly comparable to impact factor scores, which are a measure of per-article prestige. Eigenfactor is calculated based on the citations received over a five year period, excluding journal self-citations (see http://www.eigenfactor.org/). Interestingly, the Eigenfactor has resulted to be well correlated with total number of cites [36].

Webometric indicators

With the rapid diffusion of the information on the web, methods for measuring the online impact and influence of scientific information have been quickly developed giving birth to the Webometrics (or Cybermetrics) a term first coined in 1997 by T. Almind and P. Ingwersen [37].

Google Scholar is an alternative source of data to the Web of Knowledge or Elsevier’s Scopus to search citations by a page rank system based on the Autonomous Citation Index (ACI) [33]. An ACI system can automatically create a citation index from literature in electronic format by autonomously locating articles, extracting citations, identifying citations to the same article that occur in different formats, and identifying the context of citations in the body of articles. The viability of ACI depends on the ability to perform these functions accurately.

The Web Impact Factor, elaborated by P. Ingwersen in 1998, is calculated on the number
measures, including the well-known Journal
analysis is a natural choice. However, many metric
quantitative measures of scientific impact, citation
health administrators and policy makers need
measured as a function of the citations that a
influenced their work, scientific impact can be
use citations to indicate which publications
was published. Since authors of scientific papers
the impact of the journal in which the information
quality of the researcher producing it or considering
different available options are ranked following the
best available scientific evidence. Sometimes
constantly called upon to take decisions based on
public health practitioners at all levels are
How to evaluate scientific output: a few hints for
the public health practitioner
Public health practitioners at all levels are
constantly called upon to take decisions based on
the best available scientific evidence. Sometimes
different available options are ranked following the
quality of the researcher producing it or considering
the impact of the journal in which the information
was published. Since authors of scientific papers
use citations to indicate which publications
influenced their work, scientific impact can be
measured as a function of the citations that a
publication receives [40]. Therefore, when public
health administrators and policy makers need
quantitative measures of scientific impact, citation
analysis is a natural choice. However, many metric
measures, including the well-known Journal
Impact Factor, were not designed to assess the
value of individual scientists and make little sense
when applied for this purpose. Although Impact
Factor could be adjusted to fit single researcher
evaluation (adjusting by data source, specialty,
halflife of article citations, number of journals and
researchers in the field, coauthors, time periods)
[41], other measurement metrics are possibly
better suited for this task, such as the h-index.
Following the parallels proposed by M. A. Hernan,
Journal Impact Factor has biases that are evident
to trained epidemiologists (Table 1), even when
applied for its original purpose (e.g. comparing
journals) [42]. Several adjustments also seem to be
needed in the case of h-index (see Table 1).
The rigid application of quantitative metrics
to judge the quality of a journal, of a single
publication or of a researcher suffers from many
negative issues and is prone to many reasonable
criticisms. A set of editorials, recently published
in Epidemiology ([42] and references therein) and
in Nature ([7, 13] and related commentaries) have
highlighted the pros and cons of ranking research.
All the indicators are potentially influenced by
the citation behaviour of the researchers that
can be distorted due to the fact that they are
under pressure from the “Publish or perish”
dogma. As noticed by others [6], this process
can lead to writing articles with “the least
publishable unit” (salami science [43]), clearly
wrong papers, petty trivia, and ultra thin salami
slices where even their authors (except for
incurable megalomaniacs) would confidentially
acknowledge triviality (“junk” in the words
of J. P. Ioannidis and colleagues [6]). Virtually
anything can be published, as the current system
does not penalize junk publication but it rewards
productivity [6]. On the other hand, the adoption
of qualitative systems alone (e.g. peer review or
review panels) could lead to distorted choices
because of the “old boys network” bias when well
established scientists act as gatekeepers blocking
the “entrance” of new researchers who are not
already part of their network or who do not share
the same scientific views (see for example [44]).
A reasonable way forward would probably be
the use of a qualitative assessment by a panel
review founded on assessment through few
robust quantitative metrics. In any case, any
method will be imperfect if intellectual honesty
and scientific ethics are not put to the forefront.
We conclude with a sentence by BG Charlton:
“Yet real science must be an arena where truth
is the rule; or else the activity simply stops being
science and becomes something else: Zombie
science” [45].
Table 1. Epidemiological parallels of two popular metrics to assess journals and single researcher quality, Journal Impact Factor (modified from [24]) and Author h-Index.

<table>
<thead>
<tr>
<th>Hypothetical cancer follow up study</th>
<th>Journal Impact Factor</th>
<th>Author h-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>Individuals</td>
<td>Published items</td>
</tr>
<tr>
<td>Source population</td>
<td>Selected cohort</td>
<td>Journals present into ISI database</td>
</tr>
<tr>
<td>Eligibility criteria</td>
<td>Defined in the study protocol</td>
<td>Publications in the 2 years before year Y</td>
</tr>
<tr>
<td>Outcome</td>
<td>Number of cases of diseases</td>
<td>IF of journal X in year Y</td>
</tr>
<tr>
<td>Classification bias</td>
<td>Outcome misclassification</td>
<td>Citations in Journal X of items published in the same Journal X</td>
</tr>
<tr>
<td>Subjects contributing to denominator</td>
<td>Individuals at risk of developing the disease</td>
<td>Articles designed as primary, review or “front matter” by Thomson Scientific</td>
</tr>
<tr>
<td>Subjects not contributing to denominator</td>
<td>Individuals with disease</td>
<td>Commentaries, letters, editorials not considered substantive papers by Thomson Scientific</td>
</tr>
<tr>
<td>Criteria to determine which subjects contribute to the denominator</td>
<td>Defined by the study protocol</td>
<td>Unspecified</td>
</tr>
<tr>
<td>Other possible adjustment factors</td>
<td>Age, gender, race, etc.</td>
<td>Subject area</td>
</tr>
</tbody>
</table>

References
9) Nalimov VV, Mul’chenko ZM. Measurement of Science.


19) Pearce WJ. Citations: Too many, or not enough?. Scientist 2010; 24:29.


23) Rossner M, Van Epps, Hill E. Show me the data. JEM 2007;3052-3.


29) Finch A. Can We Do Better Than Existing Author Citation Metrics? Bioessays 2010;32:744-7.


